



ADVANCES IN DIGITAL SCHOLARLY EDITING



ADVANCES IN DIGITAL SCHOLARLY EDITING

PAPERS PRESENTED AT THE DIXIT CONFERENCES
IN THE HAGUE, COLOGNE, AND ANTWERP

edited by

PETER BOOT
ANNA CAPPELLOTTO
WOUT DILLEN
FRANZ FISCHER
AODHÁN KELLY
ANDREAS MERTGENS
ANNA-MARIA SICHANI
ELENA SPADINI
DIRK VAN HULLE

© 2017 Individual authors

Published by Sidestone Press, Leiden
www.sidestone.com

Imprint: Sidestone Press

Lay-out & cover design: Sidestone Press
Cover illustration: Tessa Gengnagel

ISBN 978-90-8890-483-7 (softcover)
ISBN 978-90-8890-484-4 (hardcover)
ISBN 978-90-8890-485-1 (PDF e-book)

Contents

Welcome	11
Preface	13
Introduction	15
Peter Boot, Franz Fischer & Dirk Van Hulle	
 WP1 CONCEPTS, THEORY, PRACTICE	
Towards a TEI model for the encoding of diplomatic charters. The charters of the County of Luna at the end of the Middle Ages	25
Francisco Javier Álvarez Carbajal	
The uncommon literary draft and its editorial representation	31
Mateusz Antoniuk	
Data vs. presentation. What is the core of a scholarly digital edition?	37
Gioele Barabucci, Elena Spadini & Magdalena Turska	
The formalization of textual criticism. Bridging the gap between automated collation and edited critical texts	47
Gioele Barabucci & Franz Fischer	
Modelling process and the process of modelling: the genesis of a modern literary text	55
Elli Bleeker	
Towards open, multi-source, and multi-authors digital scholarly editions. The Ampère platform	63
Christine Blondel & Marco Segala	
Accidental editors and the crowd	69
Ben Brumfield	
Toward a new realism for digital textuality	85
Fabio Ciotti	
Modelling textuality: a material culture framework	91
Arianna Ciula	
Multimodal literacies and continuous data publishing. Une question de rythme	99
Claire Clivaz	

Theorizing a digital scholarly edition of <i>Paradise Lost</i>	105
Richard Cunningham	
The digital libraries of James Joyce and Samuel Beckett	109
Tom De Keyser, Vincent Neyt, Mark Nixon & Dirk Van Hulle	
Editing the medical recipes in the Glasgow University Library Ferguson Collection	115
Isabel de la Cruz-Cabanillas	
The archival impulse and the editorial impulse	121
Paul Eggert	
Pessoa's editorial projects and publications. The digital edition as a multiple form of textual criticism	125
Ulrike Henny-Krahmer & Pedro Sepúlveda	
Reproducible editions	135
Alex Speed Kjeldsen	
'... but what should I put in a digital apparatus?' A not-so-obvious choice. New types of digital scholarly editions	141
Raffaella Afferni, Alice Borgna, Maurizio Lana, Paolo Monella & Timothy Tambassi	
Critical editions and the digital medium	145
Caroline Macé	
Scholarly editions of three rabbinic texts – one critical and two digital	149
Chaim Milikowsky	
From manuscript to digital edition. The challenges of editing early English alchemical texts	159
Sara Norja	
Towards a digital edition of the Minor Greek Geographers	165
Chiara Palladino	
Digital editions and materiality. A media-specific analysis of the first and the last edition of Michael Joyce's <i>Afternoon</i>	171
Mehdy Sedaghat Payam	
Challenges of a digital approach. Considerations for an edition of Pedro Homem de Mello's poetry	177
Elsa Pereira	
The born digital record of the writing process. A hands-on workshop on digital forensics, concepts of the forensic record and challenges of its representation in the DSE	183
Thorsten Ries	

Enduring distinctions in textual studies Peter Shillingsburg	187
Blind spots of digital editions. The case of huge text corpora in philosophy, theology and the history of sciences Andreas Speer	191
Data driven editing: materials, product and analysis Linda Spinazzè, Richard Hadden & Misha Broughton	201
Making copies Kathryn Sutherland	213
The Videotext project. Solutions for the new age of digital genetic reading Georgy Vekshin & Ekaterina Khomyakova	219
A stemmatological approach in editing the Greek New Testament. The Coherence-Based Genealogical Method Klaus Wachtel	223
WP2 TECHNOLOGY, STANDARDS, SOFTWARE	
What we talk about when we talk about collation Tara L. Andrews	231
The growing pains of an Indic epigraphic corpus Dániel Balogh	235
The challenges of automated collation of manuscripts Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Vincent Neyt & Dirk Van Hulle	241
The role of digital scholarly editors in the design of components for cooperative philology Federico Boschetti, Riccardo Del Gratta & Angelo Maria Del Grosso	249
Inventorying, transcribing, collating. Basic components of a virtual platform for scholarly editing, developed for the Historical-Critical Schnitzler Edition Stefan Büdenbender	255
Combining topic modeling and fuzzy matching techniques to build bridges between primary and secondary source materials. A test case from the King James Version Bible Mathias Coeckelbergs, Seth van Hooland & Pierre Van Hecke	261

The importance of being... object-oriented. Old means for new perspectives in digital textual scholarship	269
Angelo Mario Del Grosso, Emiliano Giovannetti & Simone Marchi	
Edition Visualization Technology 2.0. Affordable DSE publishing, support for critical editions, and more	275
Chiara Di Pietro & Roberto Rosselli Del Turco	
Compilation, transcription, multi-level annotation and gender-oriented analysis of a historical text corpus. Early Modern Ducal Correspondences in Central Germany	283
Vera Faßhauer	
<i>Hybrid scholarly edition</i> and the visualization of textual variants	289
Jiří Flaišman, Michal Kosák & Jakub Říha	
Burckhardtsource.org: where scholarly edition and semantic digital library meet	293
Costanza Giannaccini	
EVI-linhd, a virtual research environment for digital scholarly editing	301
Elena González-Blanco, Gimena del Rio, Juan José Escribano, Clara I. Martínez Cantón & Álvaro del Olmo	
Critical diplomatic editing. Applying text-critical principles as algorithms	305
Charles Li	
St-G and DIN 16518, or: requirements on type classification in the Stefan George edition	311
Frederike Neuber	
Visualizing collation results	317
Elisa Nury	
The Hebrew Bible as data: text and annotations	323
Dirk Roorda & Wido van Peursen	
Full Dublin-Core Jacket. The constraints and rewards of managing a growing collection of sources on omeka.net	333
Felicia Roşu	
Of general and homemade encoding problems	341
Daniela Schulz	
The role of the base manuscript in the collation of medieval texts	345
Elena Spadini	

A tailored approach to digitally access and prepare the 1740 Dutch <i>Resolutions of the States General</i>	351
Tuomo Toljamo	
Editorial tools and their development as a mode of mediated interaction	357
Tuomo Toljamo	
TEI Simple Processing Model. Abstraction layer for XML processing	361
Magdalena Turska	
 WP3 ACADEMIA, CULTURAL HERITAGE, SOCIETY	
Edvard Munch's Writings. Experiences from digitising the museum	367
Hilde Bøe	
Crowdfunding the digital scholarly edition. Webcomics, tip jars, and a bowl of potato salad	375
Misha Broughton	
Editing medieval charters in the digital age	383
Jan W. J. Burgers	
Editing copyrighted materials. On sharing what you can	391
Wout Dillen	
What you c(apture) is what you get. Authenticity and quality control in digitization practices	397
Wout Dillen	
The journal al-Muqtabas between Shamela.ws, HathiTrust, and GitHub. Producing open, collaborative, and fully-referencable digital editions of early Arabic periodicals – with almost no funds	401
Till Grallert	
Digital editions of artists' writings. First Van Gogh, then Mondrian	407
Leo Jansen	
Digital editing: valorisation and diverse audiences	415
Aodhán Kelly	
Social responsibilities in digital editing – DiXiT panel. Editing and society: cultural considerations for construction, dissemination and preservation of editions	421
Aodhán Kelly	
Documenting the digital edition on film	427
Merisa Martinez	

Towards a definition of ‘the social’ in knowledge work	433
Daniel Powell	
Beyond Open Access. (Re)use, impact and the ethos of openness in digital editing	439
Anna-Maria Sichani	
The business logic of digital scholarly editing and the economics of scholarly publishing	449
Anna-Maria Sichani	
The social edition in the context of open social scholarship. The case of the Devonshire Manuscript (BL Add Ms 17, 492)	453
Ray Siemens	
Nowa Panorama Literatury Polskiej (New Panorama of Polish Literature). How to present knowledge in the internet (Polish specifics of the issue)	463
Bartłomiej Szleszyński	
Digital Rockaby	467
Katerina Michalopoulou & Antonis Touloumis	

Welcome

Undoubtedly, the digital turn has challenged the theoretical understanding of and the methodological approach to the core research activity in most of the humanities. But while the Digital Humanities have reshaped substantially scholarship, there is still hardly any academic institution which is able to provide the infrastructure and the resources needed in order to train the next generation of young scholars and researchers in the diverse range of skills and methods required. For this reason, ten European leading institutions from academia, in close collaboration with the private sector and cultural heritage institutions and funded under the EU's Marie Skłodowska-Curie Actions, established one of the most innovative training networks for a new generation of scholars in the field of digital scholarly editing.

One of the key elements of the DiXiT research and training programme has been a set of three conferences, held in The Hague (2015), Cologne and Antwerp (both 2016), where external researchers joined the young researchers of the DiXiT network to discuss the continued development of digital scholarly editions. The extended abstracts and articles collected in this volume reflect the dynamics in the field of Digital Humanities which, more so than other fields of the humanities, is driven by technical solutions, their methodological adaptations and the involved research practice in both the academic and the non-academic sector.

As the coordinator of the DiXiT network, I would like to take the opportunity to express my thanks to all who have made these most prolific years of always trustful cooperation possible: the universities and institutions that formed the body of partners, my colleagues who served as supervisors in the most professional and supportive way, our fellows who contributed their stimulating ideas and projects and, over the years, formed a wonderful body of young researches, and finally the European Union and its research council that set up the MSCA-ITN-programme and supported our network in the most collaborative and generous way.

Our experience with DiXiT is a strong argument for the strength and the need of international cooperation that does not only provide the resources of a network but also shapes our way of thinking. DiXiT clearly is a European project and it could not have happened otherwise.

Andreas Speer (Coordinator of DiXiT)

Preface

The DiXiT Network ran a most successful course of attracting, stimulating, and training Early Stage and Experienced Researchers in Textual Scholarship and Digital Humanities across Europe. Logistically, the network was pivoted on individual training at widely dispersed host institutions, providing environments congenial to each DiXiT fellow's own explorative research work. The representatives of the diverse host institutions formed the DiXiT collegium. They guaranteed and facilitated its networking in manifold patterns of cooperation, culminating in a succession of three DiXiT Conventions. It was at these conventions that the Early Stage and Experienced Researchers could and did show their capacity to network, that is to aggregate collectively together the innovative gain of the Scholarly Training Programme that supported them and their work. The extended abstracts of the convention contributions here assembled showcase the variety of subjects dealt with in and around the topics of digital editing: from issues of sustainability to changes in publications cultures, from the integrity of research and intellectual rights to mixed methods applied to digital editing – to name only a few. The conventions acted importantly, too, as fruitful occasions to make such a variety of scholarly topics resonate at the social level in the cities which hosted these DiXiT events and where activities were embedded thoughtfully across institutional networks and platforms beyond academia. The DiXiT network thus has achieved a fusion of intellects and personalities strong enough to engender veritable advances in scholarly practice and thought, of which the collocated abstracts give rich evidence.

The DiXiT External Experts Advisory Board

Arianna Ciula, Gregory Crane, Hans Walter Gabler and Espen Ore

Introduction

Peter Boot,¹ Franz Fischer²

& Dirk Van Hulle³

As the papers in this volume testify, digital scholarly editing is a vibrant practice. Since high quality digital scholarly editions have been around for two decades, it is no longer a new undertaking. In fact, digital scholarly editing represents one of the longest traditions in the field of Digital Humanities – and the theories, concepts, and practices that were designed for editing in a digital environment in turn have influenced deeply the development of Digital Humanities as a discipline.⁴ That the field is still experimental in many respects is mainly because the possibilities of digital technologies are in constant flux, and ever expanding. By bringing together the extended abstracts from three conferences organised within the DiXiT project (2013-2017), this volume shows how digital scholarly editing is still developing and constantly redefining itself. So who better to answer the question of what digital scholarly editing entails than a broad selection of the community of scholars who practice it? The common denominator for these three conferences, DiXiT (Digital Scholarly Editions Initial Training Network), is a project funded under the EU's Marie Skłodowska-Curie schemes for researcher training and mobility. The conferences, held in The Hague (2015), Cologne and Antwerp (both 2016) brought together young researchers employed within the DiXiT network with external researchers to discuss the continuing development of digital scholarly editing.

Digital scholarly editions initial training network

Scholarly editing has a long-standing tradition in the humanities. It is of crucial importance within disciplines such as literary studies, philology, history, philosophy, library and information science and bibliography. Scholarly editors were among the first within the humanities to realize the potential of digital media

1 peter.boot@huygens.knaw.nl.

2 franz.fischer@uni-koeln.de.

3 dirk.vanhulle@ua.ac.be.

4 For the state of the discourse see: Driscoll and Pierazzo (2016), Pierazzo (2015), Apollon *et al.* (2014), Sahle (2013).

for performing research, for disseminating their results, and for bringing research communities together. Hence, digital scholarly editing is one of the most mature fields within the Digital Humanities. Yet, it is also confronted with many challenges. As a cross-disciplinary field, its practitioners must acquire many competences. But since technologies and standards keep improving at a rapid pace, stable practices are hard to establish. Scholars need both high skills and deep knowledge to face present and future challenges of digital scholarly editing. Before DiXiT there was no dedicated postgraduate programme able to provide the training to form a new generation of those scholars.

The institutions cooperating within the DiXiT network are some of the most respected university groups and academies working within the field of digital scholarly editing, along with partner institutions from the commercial and cultural heritage sectors. Together they represent a wide variety of technologies and approaches to European digital scholarly editing. They have created a robust research and training programme in the core skills of digital scholarly editing, reaching researchers from some 300 European institutions. Some 800 researchers in Europe and beyond participated in about 25 events.⁵ In applying for a Marie Skłodowska-Curie Initial Training Network in 2012 the network formulated a number of priorities:

1. Attracting young scholars that are able to develop the mix of competences from the humanities, computer science, and information studies that digital scholarly editions require.
2. Combining resources of the most prominent institutions and enterprises in the field in order to train these scholars in the best possible way.
3. Collaborating seamlessly across international scholarly communities in an increasingly cross-cultural and interdisciplinary field.
4. Intensifying efforts towards standards, interoperability and the accumulation of shared methods.
5. Creating suitable models and core curricula for digital scholarly editing.
6. Improving the publishing workflows, infrastructures and publishing venues for digital scholarly editions.
7. Developing a sustainable infrastructure for improving long-term prospects of digital scholarly editing projects.

Fortunately, the EU commission decided to award the requested grant to the DiXiT network. Since 2013, twelve early stage researchers and five experienced researchers are or have been employed within the network, one or two at each participating institution. Each of the early stage researchers works on his or her thesis (or thesis-size subject), while the experienced researchers have more focussed and shorter appointments. One of the unique characteristics of the Marie Curie

⁵ Participating researchers came from almost all European countries and many non-European countries; DiXiT training events and conferences were held across 11 European countries including two online summer schools in Spanish reaching out to Latin America; see <http://dixit.uni-koeln.de/programme>.

scheme is that researchers cannot apply for jobs in their own countries. Although working in different European countries, they therefore are motivated to reach out to each other. This has created a closely-knit group of researchers that works together very well.

This collection

Apart from the informal visits, secondments, research stays and a multitude of video sessions, the DiXiT training program was organized in the form of three camps (basic training in digital scholarly editing), followed by three conventions. While the training camps were primarily targeted at the DiXiT researchers, the conventions were set up so as to create an exchange with other researchers. The topics of the conventions were based loosely on the three DiXiT work packages: Theory, Practice, Methods (WP1, Antwerp), Technology, Standards, Software (WP2, The Hague), Academia, Cultural Heritage, Society (WP3, Cologne).

This collection brings together extended abstracts from those conventions. Not all presenters chose to submit their paper for inclusion in this volume. We reorganised the papers along thematic lines: a paper presented in Antwerp that fits best within the 'Technology, Standards, Software' chapter is included under that heading.

About the papers

WP1: Theories, Practices, Methods

As digital publications are reaching a stage of maturity and scholarly editors are becoming increasingly aware of the seemingly endless possibilities of hybrid or fully digital scholarly editions, the impact of the digital medium on the field of textual scholarship has become undeniable. As a result of this 'digital turn', textual scholars are now faced with new challenges and opportunities that have called for a re-evaluation of the field's established theoretical and practical framework. To satisfy this need, DiXiT organized a conference at the University of Antwerp in association with the European Society for Textual Scholarship (ESTS), focussing on this reassessment of the theories, practices, and methods of scholarly editing in general, and of the digital scholarly edition in particular. The subject, broadly covering DiXiT work package 1, also was dealt with in papers given at the other two conferences.

The achievements shown by recent digital scholarly editions demonstrate some of the potential for innovation in the digital medium including their openness and exploratory nature. These projects have developed a wide range of editorial products. That is why a first important subject of these papers is assessing and mapping these different types of digital scholarly editions, ranging from 'digital archives' to 'knowledge sites'. This includes projects with large amounts of material as well as stabilised, authoritative readings of important works from all fields of history and human culture. The apparent distinction between these digital archives and digital editions has been questioned in the past decade (Sahle 2007,

Price 2009), leading textual scholars to argue that there is in fact no impermeable border between the two, but rather a 'continuum' (Van Hulle 2009). Armed with tools such as automated collation it is up to the reader or researcher to decide in which capacity the archive/edition is used. In his opening keynote in Antwerp, Paul Eggert (this volume) translated these ideas into a 'slider model' where the digital scholarly edition gravitates between editorial and archival impulses – thus setting the tone of the conference, which would go on to explore all gradations in between.

Of course, these different types of editions each have different editorial and technological needs, and so a second important subject in the papers concerns the architecture of digital scholarly editions providing more than simply text. The symbiotic interdependency of mass digitisation and scholarly editing does not only raise the question as to how the practice of scholarly editing can be adapted to enrich this data, but also how text/image-linkage can be employed in modelling transcription methodologies that allow for enhanced studies in palaeography and codicology, or how a digital archive of manuscripts may integrate various kinds of relevant material (including, for instance, an author's complete personal library). By trying to answer these emerging questions on how digital scholarly editions are modelled and designed, the papers seek to facilitate innovative research into the editorial process, and to acknowledge a shift in the role of the editor who is enabled to focus on the interpretive consequences of variants and versions.

These technological changes also imply that through the production of digital editions, editors may be transformed into encoders and even programmers. To accommodate these new developments we now have to rethink the kind of training, skills and knowledge the new generation of editors will need to acquire. To jumpstart this discussion, a third important group of papers critically examines the goals, functions and usability of digital scholarly editions. In the same vein, the Antwerp conference aptly was closed by Kathryn Sutherland (this volume), who drew attention to a recent reflourishing of print editions and connected this to our current digital preoccupation with 'making copies', which 'cannot simply be put down to the fact that the technology allows it – that computers are good facsimile machines; something more is going on'. Sutherland related this new preoccupation to the recent reconception of bibliography as book history and to the digital medium's capacity as a medium for replicating materialities other than its own. The document underlying the edition has been raised in status thanks to developments in the reproduction of facsimiles, alerting readers to the insecurity of certain editorial conventions and challenging the editorial model in fundamental ways.

WP2 Technology, Software, Standards

Technology is an essential ingredient for the digital scholarly edition. The universal availability of computing power makes its production possible, and the global network takes care of its dissemination. Technology, however, continues to develop and that has important consequences for digital editing. The DiXiT application mentioned a few research priorities in that direction, such as the integration of web-based editorial tools into the TEI ecosystem, an investigation into the ways

that TEI and other standards can work together, a publication architecture and the integration of analytical tools (e.g. for text analysis, stylometry and visualisation) into the digital scholarly edition. The first DiXiT convention, held in The Hague in September 2015, was devoted to the subject of technology: how to apply new technologies in editions, how to integrate new technologies in tools, and how to use them for publishing editions. This section of the volume brings together a number of papers with that subject, presented at any one of the DiXiT conventions.

Tools for creating an edition are of course an important subject for digital editors. During the conferences, some presentations focused on fully-fledged environments for digital editing created for specific projects, others presented generic environments for digital editing. Some tools focus on specific preparatory parts of the edition process, such as an analysis of page images. Other papers considered the problem at a higher level of abstraction, discussing for instance the architectural principles for an editorial environment, or the role of the editor in creating such an environment.

The most important technology to be discussed at the meetings was automated collation. This wide interest in collation also was evidenced by the high turnout for the special workshop on the subject of collation held in The Hague in November 2016. In this collection the papers look at the concept of collation, at the problems of collating modern manuscripts, the need for a base manuscript in collation and at the visualisation of collation output. Most of these papers, but not all, use the CollateX collation tool to illustrate their point.

Another subject that remains of vital importance to digital editing is publication technology. Again, the papers take a variety of approaches. Solutions range from the low-tech Omeka platform to a dedicated software platform, the Edition Visualization Technology. A very promising approach is the integration of publication information into the TEI files defining the edition.

A number of other papers discussed technological problems based on specific editions. Examples are the questions of data modelling and linguistic issues in a database of Indian epigraphy, deep (socio-) linguistic annotation in a TEI context and the very complex intertextual relations between medieval capitularies.

Another issue that is often discussed is the question of the edition-as-data versus the edition-as-interface. That too is discussed in this section (as well as under WP1). And there are other papers, not so easily brought under a single heading, such as a paper about font definition in the context of the scholarly edition, a paper about topic modelling in the context of the edition, and a general reflection about tools. Taken together, the papers in this section show the dynamic relation between the fields of scholarly editing and digital technology.

WP3: Academia, Cultural Heritage, Society

Humanities scholarship responds to, explores and brings to light our shared cultural heritage. The vast majority of texts used in scholarly editions are owned by cultural heritage organisations which are increasingly moving towards mass digitisation. Scholarly editing is part of this knowledge ecosystem and contributes to the quality of rich knowledge sites for scholars and the general public.

Papers given at the Cologne Convention provided ample evidence of the diversity and dynamic nature of approaches to editing. These approaches reach from the highest methodological standards, developed over decades of textual criticism dealing with huge text corpora in philosophy, theology and the history of sciences to non-academic communities engaging with historical documents and topics as diverse as burial records, sword fighting and spiders. All these papers and some of those presented at the conventions in Antwerp and The Hague centred on the interrelationship between academia, cultural heritage, and society and how scholarly editors can have a wider impact beyond constituencies typically served by its research.

Accordingly, as a first topic the role of museums, libraries and archives on the one hand, and academic scholars and a wider public on the other has been investigated. Papers explored how the quality of digital images and texts can be measured, ensured and improved and how information from various digitization projects can be implemented into digital critical editions. The emergence of new media and formats of cultural heritage material also require new practices, for instance when using film to document the creation of an edition, creating artworks to represent in a creative deformation of a literary work or when editing electronic literature. Digital preservation of endangered cultural heritage has been addressed as a particular challenge, especially in regions of socio-political conflicts and instability with significant constraints of hardware, software, education, network capacity and power.

Web 2.0 approaches and models of public engagement have been investigated as a second research topic. The recent explosion of social networking sites which encourage wide participation from the public opens up a completely new set of challenges and raises many unanswered questions. New editorial formats such as the 'Social Edition' are aiming at combining traditional scholarly editing practices and standards with recent developments in online social media environments. Enabled and facilitated through media wiki technology and online platforms we see a growing number of community-driven editorial projects establishing scholarly standards and methods for citizen science independently from academic paternalism.

Under a third topic scholars engaged with marketing research and experimentation in order to propose viable publication models reflecting both financial sustainability for exploitation and maintenance on the one hand, and the general interest of the scholarly community in open access on the other. With regard to increasingly image and audio oriented literacies particular attention has been drawn to the rhythm of data publication that should be brought into accordance with human thinking, talking and production of knowledge.

Conclusion

The paper book is not dead. While preparing this publication, some suggested to us that the web might be a better place for extended abstracts such as the ones that we collected here. And of course the web is a wonderful place, digital scholarly editing is thriving because of it. Still, for the focussed reading that scholarship requires,

there is nothing like paper. Conference websites are also likely to disappear, and scattered over the web, unlike paper books. We are proud to offer the world this volume as a testimony to the fecundity of the DiXiT project and the creativity of young scholars facing the challenges and opportunities of the digital realm for the scholarly edition.

References

- Apollon, D. *et al.* 2014. *Digital Critical Editions*. Urbana: University of Illinois Press.
- Driscoll, M. J. and E. Pierazzo (eds). 2016. *Digital Scholarly Editing – Theories and Practices*. Cambridge: Open Book Publishers.
- Pierazzo, E. 2015. *Digital Scholarly Editing – Theories, Models and Methods*. Farnham: Ashgate
- Price, K. 2009. 'Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?' *Digital Humanities Quarterly* 3.3. <http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html>.
- Sahle, P. 2007. 'Digitales Archiv und Digitale Edition. Anmerkungen zur Begriffsklärung.' In *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*, edited by Michael Stolz. Zürich: germanistik. ch, 64-84. Online: http://www.germanistik.ch/scripts/download.php?id=Digitales_Archiv_und_digitale_Edition.
- . 2013. *Digitale Editionsformen*. SIDE 7-9 (3 vols.). Norderstedt: BoD.
- Van Hulle, D. 2009. 'Editie en/of Archief: moderne manuscripten in een digitale architectuur.' In *Verslagen en mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*, 119/2, 163-178.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° 317436.

List of DiXiT full partners

- Cologne Center for eHumanities (CCeH) – University of Cologne (UoC), Germany
- Swedish School of Library and Information Science – University of Borås (HB), Sweden
- Huygens Institute for the History of the Netherlands (Huygens ING), Department of Textual Scholarship and Literary Studies – Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), Netherlands
- Department of Digital Humanities – King's College London (KCL), United Kingdom
- Centre for Manuscript Genetics (CMG) – University of Antwerp (UA), Belgium

Center for Information Modelling in the Humanities – Graz University (GU),
Austria

An Foras Feasa – National University of Ireland, Maynooth (NUIM), Ireland

Pôle de Lyon – École des Haute Études en Sciences Sociales (EHESS), France

DigiLab – Università di Roma ‘La Sapienza’ (R1), Italy

IT Services (OXIT) – University of Oxford (UOX), United Kingdom

List of DiXiT fellows and affiliated institutions

Early Stage Researchers

Richard Hadden – An Foras Feasa, National University of Ireland, Maynooth

Tuomo Toljamo – Department of Digital Humanities, King’s College London

Elli Bleeker – Centre for Manuscript Genetics (CMG), University of Antwerp

Frederike Neuber – Center for Information Modelling in the Humanities, Graz
University

Francisco Javier Álvarez Carbajal – Pôle de Lyon, École des Haute Études en
Sciences Sociales

Elena Spadini – Huygens Institute for the History of the Netherlands (KNAW)

Misha Broughton – Cologne Center for eHumanities (CCeH), University of
Cologne

Merisa Martinez – Swedish School of Library and Information Science), University
of Borås

Daniel Powell – Department of Digital Humanities, King’s College London

Anna-Maria Sichani – Huygens Institute for the History of the Netherlands
(KNAW)

Aodhán Kelly – Centre for Manuscript Genetics (CMG), University of Antwerp

Federico Caria – DigiLab – Università di Roma ‘La Sapienza’

Experienced Researchers

Wout Dillen – Swedish School of Library and Information Science, University of
Borås

Linda Spinazzè – An Foras Feasa, National University of Ireland, Maynooth

Magdalena Turska – IT Services (OXIT), University of Oxford

Gioele Barabucci – Cologne Center for eHumanities (CCeH), University of
Cologne

Roman Bleier – Center for Information Modelling in the Humanities, Graz
University

WP1

Concepts, Theory, Practice

Towards a TEI model for the encoding of diplomatic charters

The charters of the County of Luna at the end of the Middle Ages

Francisco Javier Álvarez Carbajal¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

This paper presents the current state of my research initiated during my secondment in the Centre for Information Modelling of the Austrian Centre for Digital Humanities in Graz. My main goal was to recover the original aim of the CEI², that is, explore the idea of proposing a standard for the encoding of diplomatic sources and, eventually, find a way to fit this standard in the TEI guidelines. For this purpose, the methodology I have adopted is a case study: a prototype of a Digital Diplomatic Edition presenting an ODD designed to deal with the encoding of the diplomatic structure of a set of medieval charters. In particular, the material selected was a collection of late medieval Asturleonese charters, held in the Archivo de los *Condes de Luna*³ (León, Spain) which is an excellent sample to illustrate the documentary strategies carried out by the Quiñones family (Counts of Luna) in a period when they were engaged in a fierce rivalry with the Castilian monarchy and the urban councils of Asturias and León.⁴

1 francisco.alvarez-carbajal@ehess.fr.

2 Charters Encoding Initiative, last accessed January 27, 2016, <http://www.cei.lmu.de/index.php>.

3 The archive was catalogued several decades ago. See Álvarez Álvarez 1977 and 1982.

4 We are dedicating our doctoral research to the in-depth study of this issue.

The idea of digitizing diplomatic discourse lies on the necessity to find a feasible way to explore the documentary tenor of huge corpora of charters. This would allow to study charters at a new, larger scale, leading scholars to prove or propose hypotheses that are not achievable easily by traditional means. Besides, there is always the thorny problem of finding an encoding standard, particularly now that TEI has been raised as the *de facto* one. If digital diplomatists do not find a way to standardize their encoding, they will find their editions unable to interoperate and exchange data.

Finally, it is necessary to clarify that by ‘diplomatic edition’ I do not mean the representation of the visual aspects of manuscripts⁵. Instead, I am using ‘diplomatic’ here to refer to the scholar field of Diplomatics⁶. Therefore, my goal was to create a digital edition that assisted diplomatists to answer, test or suggest hypotheses in their field of knowledge. In order to do so, I believe that a digital diplomatic edition must implement at least the two following key elements:

- Accessible files containing the encoding of the diplomatic structure of a corpus of medieval charters. Such encoding should be as TEI compliant as possible and, in this case, aim to promote further discussion on the development of a standard.
- A search engine capable of retrieving with ease the aforementioned diplomatic clauses.

Diplomatics and digital editions

Other scholars before have reflected on the possibilities that Digital Editions offered to Diplomatics⁷. Michele Ansani, for example, stated:

*‘(Diplomatics) has widely and long time ago consolidated its methods, during decades it has discussed and substantially normalized its editorial criteria, but above all it has detailed its own analytic categories, formalized terminological and conceptual convergences, and systematized its international scholarly language. (Therefore diplomatics) is a field in theory ready to confront this (digital) transition without major trauma.’*⁸

5 This is the most extended philologic understanding of diplomatic (or documentary) editions. For instance, see Pierazzo 2011: ‘According to its classic definition, a diplomatic edition comprises a transcription that reproduces as many characteristics of the transcribed document (the diploma) as allowed by the characters used in modern print. It includes features like line breaks, page breaks, abbreviations and differentiated letter shapes.’

6 « La Diplomatique est la science qui étudie la tradition, la forme et l’élaboration des actes écrits. Son objet est d’en faire la critique, de juger de leur sincérité, d’apprécier la qualité de leur texte, de dégager des formules tous les éléments du contenu susceptibles d’être utilisés par l’historien, de les dater, enfin de les éditer ». Charters Encoding Initiative, Vocabulaire International de la Diplomatique, accessed January 27, 2016, <http://www.cei.lmu.de/VID/VID.php?1>. For the printed version see Cárcel Ortí 1997.

7 The nature of this paper does not allow an exhaustive bibliography, but a good starting point can be found in the proceedings of the two conferences dedicated to ‘Digital Diplomatics’. Vogeler 2009; Ambrosio, Barret and Vogeler 2014.

8 Translated from the original in Italian, Ansani 2003.

However, the current scenario of diplomatic editions is quite inconsistent. The lack of encoding standards has led to a panorama of encoding fragmentation and loss of interchangeability among the created editions. In this regard, CEI has been so far the only solid attempt to design a standard for the encoding of charters. However, even if it was implemented by some editions, it never truly achieved the category of standard⁹. Quite on the contrary, many editors have developed their own model *ad hoc*, and what is worse, their encoding often is not available for the user.¹⁰ On the other hand, another group of editions do follow the TEI guidelines, but their encoding does not reflect the scholarly interests of Diplomatics.¹¹ Finally, some editions do encode the diplomatic structure of charters, although not in an exhaustive way, leaving in blank important text fragments of interest for the diplomatist.¹²

TEI and the encoding of charters

My proposal is to explore TEI and figure out how the encoding of diplomatic structure can fit in its guidelines. In the last years the rise of TEI as the *de facto* standard for textual encoding has been obvious. This, together with the fact that TEI also includes some very useful modules for the work of a digital diplomatist (such as `msDescription`, `namesDates`, `transcr`¹³) undoubtedly makes attractive the possibility of making our encoded charters TEI compliant.

However, TEI also has some crucial shortcomings for the diplomatist, since there are no modules dealing with relevant diplomatic concepts, such as diplomatic discourse, documentary tradition or means of authentication (other than seals).¹⁴

As for the encoding of diplomatic clauses, the TEI enables a way around using the `<seg>` element. It stands for arbitrary segment and it represents any segmentation of text below the 'chunk' level.¹⁵ Therefore, TEI does allow to mark up any diplomatic clause with the `<seg>` element. Then, by including the `@ana` attribute, the editor can point to somewhere in the internet where the clause already is defined. This suits very well the diplomatic analysis since, as we have seen already, the *Comission Internationale de Diplomatique* has its own vocabulary available online¹⁶, making it possible to point to a huge part of our academic jargon. According to this method, it is possible to encode, for example, the protocol of a charter in the following manner:

`<seg ana='http://www.cei.lmu.de/VID/#VID_182'>`

9 Also see Vogeler 2004.

10 See for instance *Codice Diplomatico della Lombardia Medioevale*, last accessed January 27, 2016, <http://cdlm.unipv.it/>, or also *L'edizione digitale del Liber Privilegiorum Sanctae Montis Regalis Ecclesiae* last accessed January 27, 2016, <http://vatlat3880.altervista.org/>.

11 For example, *Monumenta Germaniae Historica, Constitutiones et acta publica imperatorum et regum 1357-1378*, last accessed January 28, 2016, <http://telota.bbaw.de/constitutiones/>.

12 See ASChart, last accessed January 27, 2016, <http://www.aschart.kcl.ac.uk/index.html>.

13 <http://www.tei-c.org/Guidelines/>.

14 For a definition of these terms, see VID, specifically <http://www.cei.lmu.de/VID/VID.php?24> and <http://www.cei.lmu.de/VID/VID.php?369>.

15 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-seg.html>.

16 See this paper.

This method, however, is very inconvenient in the daily work of the editor: it is too verbose (and thus inefficient) and blind. It compels the encoder/user to previously know the number of the entry in the VID instead of using explicitly the name of the encoded clause, which increases the possibility of error in the encoding.

An ODD for the encoding of charters

In short, diplomatists need to create their own set of tags so that anyone immediately can understand and follow their encoding. Luckily, TEI Roma allows for the creation of a customizable ODD and therefore add the elements which are best suitable for our encoding.¹⁷ In this case, and again thanks to the fact that the VID is available online, it is possible to use as a base the <seg> element and define as many types of diplomatic clauses as required by referring to the definitions contained in the VID. This allows for an easier, more intuitive and explicit way of encoding charters.

However, this method also has its drawbacks. One of the main problems that arises when trying to create a standard for the encoding of charters is precisely the wide variety of charters that exists, depending on several factors, such as the historical period, its geographical provenance and the documentary type. This makes it almost impossible to foresee the complete set of elements needed to encode any kind of charter. The set of tags created for the charters of the County of Luna, for example, by force is limited by the particularities of the studied documentary source and, thus, it is highly likely that it will not suffice to encode charters from other periods of geographical areas.

Furthermore, and since the VID is used as the referential conceptual framework, the mark up created is forcefully constrained by it. The VID was intended to be a general vocabulary, so in spite of working quite well with the main clauses of the documents (at least in the case of the County of Luna) it cannot guarantee a total adequacy with all kinds of diplomatic phenomena, especially in the finest level of granularity.

This is why, in order to improve and refine the model proposed here, it is my intention to use this ODD to encode charters from other areas and periods. This could be done thanks to my participation in an Iberian network devoted to the study of Portuguese and Castilian notarial documents¹⁸, and also to my collaboration with Monasterium¹⁹.

But encoding the diplomatic structure is just one of the challenges ahead when it comes to creating a standard for digital diplomatics. Concepts such as documentary tradition or means of authentication are again non-existent in TEI, and their encoding also must be taken into account when developing a TEI compliant model for the encoding of charters.

17 <http://www.tei-c.org/Roma/>.

18 Escritura, notariado y espacio urbano en la corona de Castilla y Portugal (siglos XII-XVII) (ENCAPO).

19 <http://icar-us.eu/cooperation/online-portals/monasterium-net/>.

The prototype

Having all these things in mind, and in close collaboration with Georg Vogeler and the Centre for Information Modelling of the Austrian Center for Digital Humanities²⁰, we have launched a prototype that includes all the considerations put forward in these papers²¹. The edition allows the visualization of the above mentioned encoding. Also, it has implemented a search engine that allows to retrieve information not only by places, names and dates, but also by diplomatic clauses. In the future, a documentary-type filter may be implemented to help users to compare similar diplomatic models.

With this paper I wanted to bring this issue up for a potential community of digital diplomatists that in a close future I am sure will increase as Digital Editions keep on gaining more attention from textual scholarship. The truth is that a standard always must emanate from the common agreement of the community. Probably in this regard the community of Digital Diplomatics can follow what other SIGs have achieved within TEI.

References

- Álvarez Álvarez, César, and José Ángel Martín Fuertes. 1977. *Catálogo del Archivo de los Condes de Luna*. León: Colegio Universitario de León.
- . 1982. 'Addenda al catálogo del archivo de los *Condes de Luna*.' *Archivos leoneses* 36 (71): 159-186.
- Ambrosio, Antonella, Sébastien Barret and Georg Vogeler (eds). 2014. *Digital Diplomatics. The Computer as a Tool for the Diplomatist?* Köln et al.: Böhlau.
- Ansani, Michele. 2003. 'Diplomatica e nuove tecnologie. La tradizione disciplinare fra innovazione e nemesi digitale.' *Scrineum-Rivista* 1: 10.
- Cárcel Ortí, Maria Milagros (ed.) 1994, 2nd edition, corrected, 1997. *Vocabulaire international de la diplomatie* (Commission internationale de diplomatie. Comité international des sciences historiques). Valencia: Universitat de Valencia.
- Pierazzo, Elena. 2011. 'A Rationale of Digital Documentary Editions.' *Literary and Linguistic Computing*: 463-477.
- Vogeler, Georg. 2004. 'Towards a standard of encoding medieval charters with XML.' Paper presented at the Digital Resources for Humanities, Newcastle upon Tyne, September 5-8, 2004.
- Vogeler, Georg (ed.). 2009. *Digitale Diplomatie. Neue Technologien in der historischen Arbeit mit Urkunden*. Köln et al; Archiv für Diplomatie, Schriftgeschichte, Siegel- und Wappenkunde, Beiheft 12.

20 <https://informationsmodellierung.uni-graz.at/en/>.

21 <http://glossa.uni-graz.at/context:decl>.

The uncommon literary draft and its editorial representation

Mateusz Antoniuk¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The 'uncommon draft' I am going to discuss was left by the Polish poet Aleksander Wat and simply is inextricable from his biography. While Wat's legacy is little-known outside Poland, that does not mean that his poetry has been ignored completely by scholars in the West (Venclova 2011; Shore 2006).

A few words on biography

Aleksander Wat was born in 1900 into a Jewish family that was assimilated heavily and secularized. He started his literary career around 1920 as an avant-garde poet writing in Polish. Wat's early poetical writings are mediocre and interesting only as a reflection of the temper of the times (they heavily are influenced by such European movements as Italian and Russian futurism and French surrealism). Much better from an artistic point of view are translations – Wat, who loved great classical Russian literature, translated the novel *The Brothers Karamazov* by Dostojevski. Early Wat is also interesting as the author of short prosaic stories written in the 1920s.

One of them, titled *The Wandering Jew*, can be described as a very ironic, perfectly constructed work of political fiction. Here is a summary of the plot: the Western world is going through a severe economic, social and cultural crisis and is threatened by a Chinese invasion. Nobody knows what to do except for a Jewish billionaire, the richest man on the planet. He knows that the only effective programme to save Western civilization is to reject capitalism in favour of a new ideology, by which the Holy Roman Church, Communists and Jews will join forces. The billionaire's scheme is carried out: the Holy Catholic Church takes

¹ antoniuk2@interia.pl.

political control over world and proclaims theocracy; all clergymen (including the Pope himself) are Jews who converted from Judaism to Catholicism. In this global theocracy governed by the Church and Jews, society is organized in accordance with Communist ideology: no private property, no privately-run trade or industry, everything centrally controlled and planned by the Catholic-Jewish-Communist state.

The global, supranational state works smoothly to everybody's satisfaction, except for anti-Semites. Motivated by their hatred against Jews who, as you remember, turned into Catholic clergy, anti-Semites abandon Catholicism and convert to Judaism, which earlier had been rejected by Jews. In other words: anti-Semites transform themselves into Jews, because they want to fight against Jews, who have transformed themselves into Catholics. In this situation, Jews who transformed themselves into Catholics are forced to fight against anti-Semites who have now become Jews. All of this leads to a situation in which Jews are anti-Semitic and anti-Semites are Jewish.

In his short novel published in 1927 Wat creates a world that is completely crazy. Wat's novel, however, is by no means funny. Rather, it is a catastrophic grotesque written to express the writer's concern that human history has reached the highest level of absurdity in the twentieth century and that the world has lost any sense of orientation and meaning. It is strange that the author of such a grotesque, ironic and catastrophic text as *The Wandering Jew* was able to support any political idea, but in fact Wat was able. In the 1930s, Wat abandoned literary activity almost completely and became entirely involved in the Communist movement as a supporter of the Polish Communist party, which in interwar Poland was illegal.

In 1939, the Second World War began. Poland was invaded by and divided between Germany and the Soviet Union. Wat found himself in the part of Poland occupied by the Soviets, which meant that he did not perish in the Holocaust. However he became the victim of Stalin's policy. He was imprisoned and sent to a forced labour camp in Kazakhstan. But he was lucky enough to survive. When the war was over, the poet and his family were released and returned to Poland. Then, after a few years, they managed to emigrate to France.

In the post-war years, two major changes occurred in Wat's life. First, he became completely disillusioned with Communism. Second, he started to suffer from an incurable neurological illness, which caused extreme pain in the head and the face. The pain was acute and recurred frequently, with brief moments of relief. The intensity of suffering was the reason for Wat's dramatic decision: he committed suicide by taking an overdose of painkillers in 1967. Next to the unconscious poet his wife found a note that read: 'For God's sake – do not rescue'.

During this final stage of his life in the 1950s and the 1960s Wat resumed writing poetry. This time the result was impressive. The poems written during this period gained him the reputation of one of the best Polish poets of the second half of the twentieth century. The main subject of Wat's late poetic works is his traumatic experience of pain.

The poem

An untitled poem written by Wat in his final years belongs to this strand of his poetic activity. Let us look at the main fragment of the text in an English paraphrase, which is simple and inevitably loses the linguistic, stylistic values of the original (unfortunately the poem has never been translated into English by professional translator). I hope, however, that the poem will become intelligible, at least at the basic level.

*At
the peak
of antinomy:
what belongs to time
versus
what belongs to space.*

*A young man in his audacity, I declared myself the emperor of space,
and the enemy of confinement and time.
That was foolishness
of youth.
(...)
Now that I have been defeated and broken by time,
the gracious space
offers several cubic decimetres
of lifelong sentence
to hold my bones.*

*In the absence of a notary, the contract was concluded at a drinking bar.
Where the barmaid and the blacksmith signed with X's,
I, in my vanity, gave a show of calligraphy:
In a decisive move, I put an X on a piece of skin
torn off my forehead, and so,
Marsyas, no Apollo was needed to flay me,
as my forehead is useless
and aches a lot.*

Much can be said about this poem written around 1967, shortly before the author's death, but what I find most important here is the final image. A suffering man on the verge of death concludes an agreement concerning the place for his grave. The text of the agreement is written on a piece of skin torn off the man's forehead. This macabre manuscript imagined in Wat's poem is a metaphor of the conjunction between pain and writing, between body and text.

The rough draft

The rough draft of the poem is preserved in the Beinecke Rare Book and Manuscript Library, Yale University. The text was inscribed, or rather created, on a quite strange 'work sheet'. What serves here as a work sheet is a packaging of a medicine. Glifanan is a strong painkiller that Wat took to achieve temporary relief from the pain in his head and face. The written text is combined into a material unity together with the packaging of a painkiller drug – what kind of conclusion can be drawn from this strict coexistence of the scripture and its physical medium?

My answer will be as concise and simple as possible: the poem about the connection between writing and suffering not only expresses this connection with words, but also manifests it with its material form. This interpretation also can be expressed in terms proposed by American scholars studying the materiality of the text: linguistic code and bibliographic code (Bornstein 2001, 7). We can say that the key concept presented by Aleksander Wat, that is the connection between pain and writing, is communicated here simultaneously by the linguistic code and the bibliographic code. The bibliographic code underscores and enhances the meaning of the linguistic code, while the linguistic code finds its visualisation in the bibliographic code. The division between form and content cannot be maintained any longer as we are dealing with a striking unity.

What happened when the text was transferred from the manuscript to a typescript (and then to a printed edition)? What was the effect of the textual transmission process? On the one hand, we can say that the existence of the text gained a new level of intensity. In its multiple form the text is more accessible, functions in society, and can be read and interpreted. But another interpretation is also possible: in the process of textual transmission the text loses some of the intensity of its existence. It still expresses the idea of conjunction between pain and writing, but the meaning is conveyed in one dimension only: that of the linguistic code which, of course, is preserved, but not that of the bibliographic code, which has been altered completely. The 'lost value' of the transferred text can be also described using Benjamin's famous notion of 'aura' (Benjamin 1969). According to Benjamin, the aura is a special property of a work of art in its original spatial and temporal context. If an old painting is removed from the chapel where it was originally placed in accordance with the intentions of the artist and the donator, and is transferred to a museum, it loses its aura. By analogy, Wat's text moved from its original context, that is from the packaging of a medicine, loses some of its initial impact and its new presence on a paper sheet in a standard edition is much less remarkable and striking.

Editorial representation

What model of editorial representation should be employed in the case of this late, 'somatic' poem by Aleksander Wat? Maybe what this literary work needs is a genetic edition, possibly digital. It is not difficult to imagine an edition having the form of a website; the user would be able to trace the process of textual transmission in both directions: beginning with the draft, through the typescript to the first edition and the other way round. Each stage would be accompanied by extensive

commentary. Links would lead to information about Wat's illness (reflected, in such a dramatic way, by the draft) but also to information about Wat's wife, Ola, who produced the typescript and who also had a very interesting biography. Other links would point out different ways of interpreting the poem and provide the user with different contexts.

Would such an edition restore the original aura of Wat's text? According to Benjamin it would not, because any mechanical reproduction, including a digital edition, effaces the auratic properties of a work of art. But even if this conclusion is true, it is not essentially depressing. Quite the contrary, it is a consolation. Even if all manuscripts in the world were digitalized and shared online, traditional archives preserving originals, would still be needed. The packaging of Glifanan on which Wat wrote one of his most disturbing poems, would become just another speck in the global haze of virtual objects. But the original would still be waiting in Beinecke Library for a discoverer eager to see it and to touch it. Of course, this desire for physical contact with the source can be discredited easily as fetishism, logocentrism or an illusion of the metaphysics of presence. And yet many of us, I guess, still feel this kind of desire.

References

- Benjamin, Walter. 1969. *Illuminations: Essays and Reflections*, edited by Hannah Arendt. New York.
- Bornstein, George. 2001. *Material Modernism. The Politics of the Page*. New York: Cambridge University Press.
- Shore, Marci. 2009. *Caviar and Ashes: A Warsaw Generation's Life and Death in Marxism, 1918-1968*. Yale University Press.
- Venclova, Tomas. 2011. *Aleksander Wat. Life and Art of an Iconoclast*. Yale University Press.

Data vs. presentation

What is the core of a scholarly digital edition?

Gioele Barabucci,¹ Elena Spadini²

& Magdalena Turska³

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Critical editions historically have been published as printed books, they are now more often published as electronic resources, mostly as websites. The introduction of electronic publishing toolchains exposed a fundamental distinction between the *data* of the edition, usually XML/TEI files, and its *presentation*, usually HTML pages generated on the fly from the TEI files.

This distinction leads to an important question: what constitutes the core of an edition? Its data or its presentation? It is possible to think of a critical edition as a collection of pieces of pure *data*? Or is a representation layer fundamental to the concept of 'edition'?

This paper summarizes the two extreme positions that, for the sake of argument, have been held in the panel: the core of the edition lies only in the data vs. the presentation constitutes the core of the scholarly effort. The paper also reports some of the remarks that have been expressed during the discussion that followed the panel.

1 gioele.barabucci@uni-koeln.de.

2 elena.spadini@unil.ch.

3 turma@gmail.com.

A look at a different field: cartography

Before delving into the main question of the panel, it may be beneficial to have a look at another field that has gone through a similar transformation and is facing a very similar issue: cartography.

Cartography is the science, art and craft of summarizing and representing geographical information in maps, two-dimensional graphical documents. The first maps of the earth are contemporary of the first written documents (Harley and Woodward 1987; 1994; Wolodtschenko and Forner 2007).

Each map represents the knowledge that the cartographer has about a certain place plus their ability to represent that knowledge in a drawing. Advances in geographical knowledge lead to better maps, but so did also advances in drawing techniques.

In any case, advancements in cartography will never resolve a fundamental problem: that the perfect map cannot be realized. A perfectly precise depiction of the world is going to be as big as the world itself. This phenomenon is usually referred to as ‘the map is not the territory’ (Korzybski 1933) and forms the basis for the one-paragraph short-story ‘On rigor in science’ by Jorge Luis Borges.

In modern times, thus, with more high-quality data available than can be displayed, the skill of a cartographer is measured more in terms of what they decide to leave out of the map rather than what they include. In a map, each design decision carries a, sometimes unconscious, stance (Wood 1992). For example, maps of the Earth that use the Gall-Peters projection show Africa twice as big as maps that use the more common Mercator projection, conveying a substantial political message (Monmonier 1994).

Sometimes cartographers have produced maps that were ‘defective’ by design: for example nautical maps for routes around meridians and the equator use the gnomonic projection that severely distorts the dry land area but allows the nautical routes to be drawn as straight lines instead of curved lines, significantly simplifying many route calculations (Snyder 1987).

The work of cartographers has changed in the recent times with the digitalization of cartography. In particular, the crowdsourcing project OpenStreetMap has become the reference point for all digital cartographical activities since its inception in 2006.

In 2015 geographical data is no longer scarce: between satellite images and user-collected GPS tracks, an open project like OpenStreetMap⁴ has been able to amass more data than any single human will ever be able to look at or review. The problems these days are of a different sort. First of all it is very hard to edit such a gigantic amount of data when errors are spotted or updates are needed. It is also hard to make sure that the quality of the data is high enough through different subsections of the maps. Last, the concept of ‘authoriality’ is hard to apply to such a map, especially in its classic sense.

Based on the data in the OpenStreetMap’s database, many different graphical visualizations have been implemented. A few of these visualizations have been created by the project itself, but many are created and used by external public and

4 <<http://openstreetmap.org>>

private organizations that take the freely available data and render it using their own styles for their own needs.

One of the tenets of OpenStreetMap is the rule that says ‘don’t tag for the renderer’⁵. This rule highlights that the data should be correct for its own sake and no changes should be done just to make one of the many renderers generate a better visualization.

The importance of this rule stems from the fact that the *data* is the only thing that users are supposed to touch and influence directly. Their knowledge should be reflected only in what is in the database. The presentation is generated on the fly, it is thus secondary. Under these premises it would be easy to stipulate that for the OpenStreetMap project, the map is the data.

But things are never as easy as they seem. For most users the *default visualization* is the only visualization they use, know or care about. So it is understandable that they try to modify the data in order to achieve a more correct visualization when they spot something that it is not perfectly right. In some cases what moves the users are artefacts that are caused by programming errors in the render and that should be addressed by fixing the code of the render. In other cases the users perceive as *wrong* artefacts that are instead just unfamiliar (but correct) and try to modify the data so to make the visualisation ‘more right’. An example of this behaviour is the ‘cafe/bar/pub/biergarten’ problem. Each of these kinds of facility is displayed with a different icon in the default renderer. These icons are based on the official definition of those terms, definitions born in the UK culture: for example, the icon for ‘cafe’ is a coffee cup while that for ‘bar’ is a cocktail glass. There have been many cases in which people from Italy, who were not familiar with the definitions and the tagging schema, tried to change a venue from ‘cafe’ to ‘bar’ because the Italian word for that kind of place is ‘bar’, only to later complain that the map was wrong because it displayed a cocktail glass for a place in which cocktails are not sold. All these attempts to change the data are caused by the users thinking that ‘the map is wrong’ because the visualization is, or is perceived as, wrong. This should make us think that, for many users of the maps produced by OpenStreetMap, the map is the presentation.

Map	Edition
Drawing	Book
The map is the territory	The transcription is the manuscript
Database	Scanned facsimile + TEI files
Tagging schema	TEI guidelines
Renderer	HTML pages
Default render	Edition website
Change renders	Use the TEI files unchanged in another edition
The ‘cafe/bar/pub/biergarten’ problem	Presentational markup like ‘rend=‘color: red’’

5 <http://wiki.openstreetmap.org/wiki/Tagging_for_the_renderer>.

It may be sensible, at this point, to make some explicit links between maps and critical editions, so to make the analogy more explicit and more fruitful to the discussion.

This excursion into the world of cartography was supposed to abstract us by the question at stake (*what constitutes an edition: its data or its presentation?*) by providing insights from a different research area. Seeing the problem under a different light and with a more distant glance can help framing the question in a more rigorous way. The question itself, however, remains yet unsolved. In the next sections we will discuss more in depth the consequences of considering only the data or only the presentation as the core of a scholarly edition.

Plaidoyer for the presentation

There are two common places about interfaces and, as many common places, they are partly true. The first is quoted by Donald A. Norman, a design guru: ‘The real problem with interface is that it is an interface. Interfaces get in the way. I do not want to focus my energies on interface. I want to focus on my job’ (Laurel and Montford 1990). The second is that developing an interface is often the last step in a project; at that stage it may happen that the project is running out of time and money.

These widespread approaches explain why it is necessary to defend the importance of the presentation layer in a Scholarly Digital Edition.

In the Guidelines of the MLA we can read that ‘the scholarly edition’s basic task is to **present** a reliable text’ (MLA 2011; emphasis mine). In Sahle’s definition, ‘A scholarly edition is an information resource’ (Sahle 2014). The same concept is stated in the *White Paper of the MLA* devoted to Digital Editions: ‘Our definition of an edition begins with the idea that all editions are **mediations** of some kind: they are a medium through which we **encounter** some text or document and through which we can study it. In this sense an edition is a **re-presentation**, a representational apparatus, and as such it carries the responsibility not only to achieve that mediation but also to explain it: to make the apparatus visible and accessible to criticism’ (MLA 2015; emphasis mine).

By editing a text and providing a scholarly edition, the editor makes a text available. In this act of communication, as in any other, the presentation layer (the signifier) is important. To summarize, if one of the editor’s aim is to communicate, she should find out the best way to do it.

By way of illustration, some examples will be provided. A scholarly edition commonly includes a table of contents and one or more indexes. In *The Dynamic Table of Contents*⁶ (INKE *et al.*) they are presented together. The user has access to the text (on the right), to the encoding (the ‘tags’ column on the left) and to the structure of the text (the table of contents). The functionalities have been combined: when the user selects a tag, she will also see where the tag appears in the text and in the table of contents. The access to the data is in this case easy

6 <<http://voyant-tools.org/?skin=dte&corpus=1369422830248.654&docId=d1369370428939.8c168831-fad0-0412-cc3b-c8b0c0242772>>

and multiple. Therefore the act of communication undertaken by the editor is effective.

An opposite example is provided by *The Proceedings of the Old Bailey* (Hitchcock 2012), a commendable edition of criminal trials held at London's central criminal court between 1674 and 1913. The XML data (available on click) reveals that some information is encoded, but not visualized. For instance, in the trial of Sarah Crouch for a theft⁷ (*Old Bailey Proceedings Online* t17270222-67), the Three Compasses is mentioned. Three Compasses is encoded as a place, using <placeName>, but nothing in the interface shows it, nor is there a link to the beautiful map⁸ (Greenwood: Three Compasses, map section 2), from the same period, where the place is marked. The presentation layer provides in this case little access to the data, that is almost lost. The act of communication is therefore not effective.

The prototype *Around a sequence and some notes of Notebook 46: encoding issues about Proust's drafts* (Pierazzo and André 2012) shows how a peculiar document may need a peculiar presentation, and an unsuitable presentation, may even hide its interesting features. The prototype offers an interactive genetic edition; the main point of interest for the current discussion is that the visualization allows the user to look at the openings, not only at the recto or verso. As one of the editors wrote, 'Proust considered his writing space to be the opening as a whole, as he used to write only on the right side of the opening of his own notebooks and used the left side for additions, corrections and rewriting. Therefore, the page-by-page visualization that has become the standard for some types of digital editions was not an option' (Pierazzo 2015).

Visualization tools are so largely used in Digital Humanities that they can represent a distinctive mark of the discipline. The importance of a certain presentation of data is (or should be) particularly clear in the field of scholarly editions: textual scholars are used to devote attention to how, where and why a certain text is displayed on the page and in the book. Moreover, the shift from a printed book to a digital screen has stimulated reflections on how to present data.

Visualization has been used so intensively in DH that critical voices have raised. For instance, Johanna Drucker urges the Humanities to rethink visualization of data, or, as she put it, of *capta* (what the scholar collects as data). In her view, the interpretative (vs. realist) and qualitative (vs. quantitative) aspects of the information, distinctive of the Humanities, should be represented and highlighted. 'The rhetorical force of graphical display is too important a field for its design to be adopted without critical scrutiny and the full force of theoretical insight' (Drucker 2011).

'The rhetorical force of graphical display' (from quote above) is put into effect when somebody states: 'You doubt of what I say? I'll show you' (*cf.* Latour 1985). Applying this to digital editions means using 'the rhetorical force of graphical display' and (we shall add) the power of interactivity to convey the contents of the scholarly edition.

7 <<http://www.oldbaileyonline.org/browse.jsp?id=t17270222-67&div=t17270222-67>>.

8 <http://www.oldbaileyonline.org/maps.jsp?map=green&map_item_id=7506&tagtype=2&mclass=f>.

The following examples, by offering different visualizations (more or less graphical representations, more or less interactive) of the same data, show how the communication between the editor and the user may improve thanks to these approaches.

As part of the *Interface Critique* platform, J. Van Zundert and T. Andrews presented a paper entitled *Apparatus vs. Graph – an Interface as Scholarly Argument* (Van Zundert and Andrews 2014). Several visualizations of the collation results (using, for instance, CollateX) are compared: in the graph it is possible to follow the text and its fluidity for each of the witnesses. The graphical representation is readable and intuitive, while the apparatus is a technical, formalized and very compressed way to convey the information.

On the matter of interactivity, the VBase tool integrated into the *Commedia Digital Edition* (Shaw 2010) is an effective example. The readings of multiple witnesses for one verse can be visualized in the edition in a number of ways (through the apparatus or word-by-word). VBase, instead, permits the user to search for readings, according to parameters such as manuscripts and portion of the text. In this case, data is not only visualized, but the user can interact with them.

In *Information Visualization for Humanities Scholars*, the authors argue that ‘the humanities approach consists (...) of examining the objects of study from as many reasonable and original perspectives as possible to develop convincing interpretations. (...) a visualization that produces a single output for a given body of material is of limited usefulness; a visualization that provides many ways to interact with the data, viewed from different perspectives, is better; a visualization that contributes to new and emergent ways of understanding the material is best. In this context, there is an important difference between static and interactive visualizations’ (Sinclair *et al.* 2013).

The edition *L'édit de Nantes et ses antécédents* (1562-1598) by the École nationale des Chartes (Barbiche 2005-2011) is completed with three indexes: of places, persons and subjects. The *Index des lieux* lists the places and indicates in which edicts they appear. The same data (the texts of the edicts) has been processed through *RezoViz*⁹, one of the Voyant visualization tools, which shows the relations between the terms (peoples, locations, organizations). The two visualizations permit to pose different questions to the corpus: where does a certain term appear and together with which other term. The second is interactive: links can be edited and the functions at the bottom of the page set. In this case, the *Index des lieux* and the *RezoViz* visualization would be complementary, offering ‘many ways to interact with the data’.

If the majority of the SDEs offer multiple views on the material (see, e.g., Van Hulle and Neyt), ‘a visualization that contributes to new and emergent ways of understanding (it)’ is more rare in the digital editions’ panorama. To conclude, thanks to the power of graphical display and of interactivity it is possible to look at the same data in different ways and to ask them different questions. The way data

9 <<http://voyant-tools.org/tool/RezoViz/?corpus=1440929815287.8901&items=30&category=location>>.

is presented is fundamental in order to fulfil the task of any scholarly edition as an act of communication.

Supremacy of data

We have seen that digital cartography already is asking itself ‘what is the map? the data or its presentation?’. Scholarly editions in the digital era move as well from the tangible, physical object of printed editions into constantly evolving, multi-faceted and interactive applications, transferring significant portion of the power to approach the material into the hands of its users. Basically all printed editions are available only in one particular book format, while current HTML-based editions routinely are displayed in a myriad of different ways depending on the electronic device used to access them. More importantly an edition printed on paper allows only one possible reading direction while hypertextual interfaces allow the same text to be experienced in countless different ways.

Notably, for digital maps and editions alike, there is a prominent group of users who are happy to stay unconscious of the distinction between the data and the *product* they use, which is a certain projection of the data created as a response to their queries by a certain system. Nevertheless the process of transformation of the data to generate that particular view requested by the user happens every time such a request occurs. Thus, unlike the printed edition or a map, which stay safe and useful on the library shelf even if their creators switch jobs and burn all their working notes, systems that generate their digital counterparts are rendered helpless when the data is broken.

The data gathering phase is always an equally crucial prerequisite in preparing a faithful map as an edition. Alas, once the final opus is deemed ready and fixed in tangible form the work no longer depends for its existence on the original data. In digital realm this link is much more direct and can never be broken. This is the first important consequence of digital paradigm: data is not just used once in preparation and later filed away but needs to stay there forever (read: for the lifetime of the digital resource). What follows is the second consequence: changes to final presentation require edits to the underlying data or enhancements for the rendering software, but tweaking the generated output makes no sense at all.

It is unquestionable that certain presentational aspects, like system of customary visual clues adopted to communicate common phenomena (e.g. foreign words marked in italics), clear layout and user friendly interface, well-thought search engine or meaningful visualizations enormously facilitate decoding of the information stored in the raw datasets. For the majority of users it makes a world of difference when it comes to practical exploration of the scholarly resource. Yet, the change is merely quantitative, not qualitative. Even with most unhelpful of presentations, the information still stays there, in the raw data, and remains discoverable and processable in an automated manner regardless of how much the presentation layer assists in that or not.

Thus, from a general perspective, a poorly designed system with high quality data is indeed much better than a beautiful one with data that is limited, inconsistent or inadequately modelled in the digital form. For the former it will be

always possible to transfer the data into a better system, while there is not much that can be done without significant and difficult (in terms of time, money and human skills) amendments if the structure or scope of the data is insufficient.

As to what distinguishes a poor system from the well-designed one – in this day and age it sometimes will be time alone. With technological progression and growing user expectations it is often a question of mere years for the state of the art presentation to start to feel dated and obsolete. There is no reason though while such a system could not be updated at relatively small cost and given a new lease on life through changes to the presentation generating parts alone. Presentation being thus ephemeral by its very nature cannot be deemed the most important aspect of a digital work.

Final notes

The concept of data and of presentation, and the relationship between them, is changing in the digital paradigm. In the field of scholarly editions, these topics deserve new attention. The discussion that followed the panel was extensive and exposed numerous viewpoints on the subject. Unfortunately there are no transcriptions of the whole discussion (and a diligent reader will quickly observe how we thus are running into an insufficient data problem), but it can be reconstructed partly from the messages sent via Twitter with the hashtag #dixit1. An archive of all the tweets of the conference can be found on the DiXiT website.¹⁰

To conclude, we would like to point out two of the comments from the discussion. The first one is about the fact that the edition is in the encoding¹¹; this implies that encoded data is, in a certain sense, already a scholarly-mediated presentation of other data that exist in the original manuscript. The second comment reminds us that this discussion is well known and is part of a deeply entrenched philosophical argument closely related to ‘Kant’s distinction between *Erscheinung* and *Ding an sich*’.¹²

References

- Barbiche, B. (ed) 2005-2011. *L'édit de Nantes et ses antécédents (1562-1598)*. ELEC. <<http://elec.enc.sorbonne.fr/editsdepacification/>>.
- Drucker, J. 2011. ‘Humanities Approaches to Interface Theory.’, edited by Frabetti Federica. *Culture Machine* 12 (Special issue ‘The Digital Humanities, Beyond Computing’). <<http://www.culturemachine.net/index.php/cm/issue/view/23>>.
- Harley, J. B., and David Woodward (eds). 1987. *The History of Cartography, Volume 1: Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean*. University of Chicago Press.
- . 1994. *The History of Cartography, Volume 2, Book 2: Cartography in the Traditional East and Southeast Asian Societies*. University of Chicago Press.

10 <<http://dixit.uni-koeln.de/dixit1-tweetwall>>.

11 <<https://twitter.com/CCeHum/status/644498034479509504>>.

12 <<https://twitter.com/PeursenWTVan/status/644504768329740288>>.

- Hitchcock, Tim, Robert Shoemaker, Clive Emsley, Sharon Howard and Jamie McLaughlin, *The Old Bailey Proceedings Online, 1674-1913*. version 7.0, 24 March 2012 <<http://www.oldbaileyonline.org>>.
- INKE, Canadian Writing Research Collaboratory, UAP Virginia Tech, and Voyant Tools. 'Dynamic Table of Contexts.' <<http://inke.ca/projects/tools-and-prototypes/>>
- Korzybski, A. 1933. *Science and Sanity. An Introduction to Non-Aristotelian Systems and General Semantics*. Oxford.
- Latour, B. 1985. 'Visualisation and Cognition: Drawing Things Together.' In *Knowledge and Society: Studies in the Sociology of Culture Past and Present : A Research Annual*, edited by H. Kuklick, 1-40. Greenwich, Conn.: Jai Press. <<http://worrydream.com/refs/Latour%20-%20Visualisation%20and%20Cognition.pdf>>.
- Laurel, B, and S. Joy Mountford (eds). 1990. *The Art of Human-Computer Interface Design*. Reading, Mass.: Addison-Wesley Pub. Co.
- MLA. 2011. 'Guidelines for Editors of Scholarly Editions.' <<https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/Guidelines-for-Editors-of-Scholarly-Editions>>.
- MLA. 2015. 'Considering the Scholarly Edition in the Digital Age: A White Paper of the Modern Language Association's Committee on Scholarly Editions.' <<https://scholarlyeditions.commons.mla.org/2015/09/02/cse-white-paper/>>.
- Monmonier, M. 1994. *Drawing the Line: Tales of Maps and Cartocontroversy. 1st edition*. New York: Henry Holt & Co.
- Pierazzo, E., and J. André. 2012. 'Autour D'une Séquence et Des Notes Du Cahier 46: Enjeu Du Codage Dans Les Brouillons de Proust.' <http://research.cch.kcl.ac.uk/proust_prototype/index.html>.
- Pierazzo, E. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey ; Burlington, VT: Ashgate.
- Sahle, P. 2014. 'Criteria for Reviewing Scholarly Digital Editions, Version 1.1' <<http://www.i-d-e.de/publikationen/weitereschriften/criteria-version-1-1/>>.
- Shaw, P. (ed.) 2010. *Dante Alighieri Commedia A Digital Edition*. SDE – SISMELE <<http://sd-editions.com/AnaAdditional/commediaonline/home.html>>.
- Sinclair, S., S. Ruecker, and M. Radzikowska. 2013. 'Information Visualization for Humanities Scholars.' In *Literary Studies in the Digital Age*, edited by Kenneth M. Price and Ray Siemens. Modern Language Association of America. <<http://dlsanthology.commons.mla.org/information-visualization-for-humanities-scholars/>>.
- Snyder, J.P. 1987. *Map Projections--a Working Manual*. U. S. Government Printing Office.
- Van Hulle, D., and V. Neyt (eds). 2011. *The Beckett Digital Manuscript Project*. Brussels: University Press Antwerp (ASP/UPA). <<http://www.beckettarchive.org>>.

- Van Zundert, J., and T. Andrews. 2014. *Apparatus vs. Graph – an Interface as Scholarly Argument*. Interface Critique. <<https://vimeo.com/114242362>>.
- Wolodtschenko, A., and T. Forner. 2007. 'Prehistoric and Early Historic Maps in Europe: Conception of Cd-Atlas.' *E-Perimetron* 2 (2): 114-16.
- Wood, D. 1992. *The Power of Maps*. The Guilford Press.

The formalization of textual criticism

Bridging the gap between automated collation and edited critical texts

Gioele Barabucci¹ & Franz Fischer²

Paper presented at 'Digital scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Introduction

Not much of the editorial process of creating digital scholarly editions is automated or assisted by computer tools. Even those parts of the process that are automated do not really form a cohesive unit. For example, the tools and the data formats used to extract text via OCR from manuscripts are different and mostly incompatible with the tools and data formats used to collate the extracted texts. Because of these incompatibilities or missing computer tools, the editorial workflow is highly fragmented, hard to manage in small-scale editions and to extend to big-scale projects.

This paper discusses several challenges we have been facing in various collaborative editorial projects run at our research center, including large-scale projects, such as *Capitularia* or *Novum Testamentum Graece*.³ These editions deal with hundreds texts transmitted in even more manuscript witnesses. Conceived as long-term project, they involve dozens of editors and collaborators. It is easy to imagine that maintaining consistent editorial practices throughout the years and changes in the composition of the editing team is not an easy effort. Simple tasks like collating a particular passage or finding occurrences of a certain editing pattern in the transmission of the texts can be daunting. There are computer tools that could help with some of these tasks. However, they cover only a few of the editorial steps, do not always interact with the editors and are hard to combine into a coherent workflow.

1 gioele.barabucci@uni-koeln.de.

2 franz.fischer@uni-koeln.de.

3 See the contributions by Klaus Wachtel and Daniela Schulz (this volume).

In order to solve these problems, we suggest the adoption of a shared formalization describing the editorial process. The use of this shared formalization will allow the whole editorial process to be semi-automated, with positive repercussions on the workload of the editors and on the quality and verifiability of the edition itself.

Problem: computers should help more but there are many gaps in the editorial workflow

There are many areas of the editorial process that could be improved if computer-based tools were available.

Dealing with massive traditions. Editorial projects that deal with hundreds of witnesses often have to sacrifice precision in their results in order to be able to deliver a complete edition inside the limits of the agreed budget (time, people and money). Letting computers deal with the most repetitive tasks frees up many resources that can be better spent on the research questions of the projects.

Advanced search. The current search tools allow only few kinds of searches, usually just textual searches. Researchers often have the need to search vast corpora looking for complex editing patterns.

Documentation of editorial guidelines and automatic review. Normally the editorial guidelines (e.g., which variants are to be included in the critical apparatus) are expressed in the introduction of the edition. It is impossible in practice for the readers of the editions, as well as for the authors themselves, to be sure that these guidelines have always been followed.

Reproducibility. In theory, given the same materials and the rules stated in the introduction of an edition, it should be possible to reproduce the same outcomes described in the edition itself. This is what gives credibility to an edition. (For a discussion on the role of reproducibility in digital scholarly editions see, among others, van Zundert 2016.) Such verification tasks are impossible to carry out manually for any non-trivial edition.

Admittedly, certain parts of the editorial process have received a certain degree of support from computer tools. For example, collation tools such as CollateX (Dekker 2015) have been successfully integrated in many editions; stemmatic tools like Stemmaweb (Andrews 2014) also have been applied in various projects; publication frameworks based on TEI and LaTeX like TEI-CAT (Burghart 2016) or EVT (Rosselli Del Turco 2014) have been used in the production of some editions.

All these tools, however, act like small unconnected islands. They expect input and output data to match their own data format and data model, both narrowly tailored to their task and following their own idiosyncratic vocabulary. Please note that, while this behaviour seems to resembles the famous UNIX principle ‘do one thing and do it well’ (Salus 1994), it fails to comply with the second, and more important, UNIX principle ‘write programs to work together’. In our experience with both small- and large-scale projects, most of the programming time is spent writing code to coordinate these services (for example converting between incompatible data models). Instead of writing glue code between incompatible services, the same time could be better spent providing enhanced functionality built *on top* of these tools.

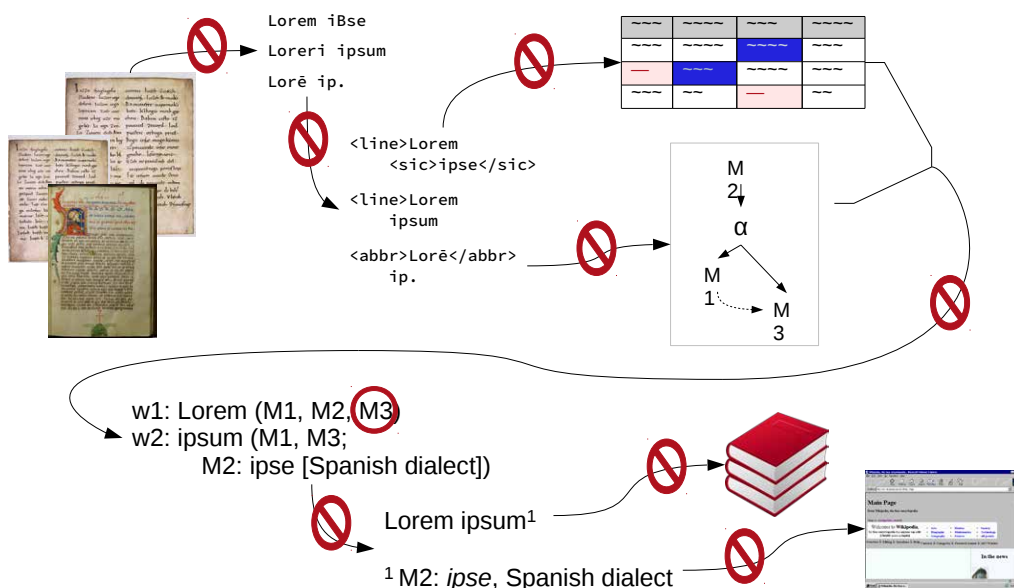


Figure 1: The problem: current tools do not communicate between each other.

Looking at a generic editorial process for a digital scholarly edition (see Figure 1), we easily notice that there is a roadblock between each step. Each of these roadblocks represent a different data format, data model or vocabulary. At each step some editorial information is lost because of these incompatibilities. Moreover, the inherent difficulty in dealing with all these disconnected worlds leads editors to perform many of these steps manually.

Whenever editors perform a step manually, various issues arise. First of all, doing manual transformations often means manually changing files *in situ*, mixing the output of the tool with the manual interventions of the editor. Second, manual changes are hard (often impossible) to replicate. This means that if one of the steps has to be redone (for example, because a better transcription has been produced or a new witness has emerged), then the editors will not be able to use the computer tool again and will have to redo the whole step manually, skip it or lose their interventions.

We can provide a practical example of how these manual changes interfere with the editing process. Suppose we are editing a three-witness text. We used CollateX to generate a collation table of our three witnesses. Because there were some misalignment, we manually fixed the generated collation. We then used this collation to manually typeset an edition in CTE, manually choosing some readings as correct and including in the critical apparatus all the variants except some deemed irrelevant (such as all orthographic variants except in proper nouns). In addition, we decided to render all the names of kings in bold. We realize only near the end of our edition that fragments of our text can be found also in fourth manuscripts. Generating a new collation table with CollateX means losing all the manual work we have done in the meantime, basically the whole editorial work. If, instead of manually making these changes, we just stated the changes we wanted

to make and let a computer apply them for us, we could easily run the whole editorial process again in few clicks. As a useful byproduct, we also would have a complete list of all the editorial decisions we have taken during the preparation of the edition. But how could we describe the actions we want to make?

Root cause: current tools are based on incomplete theoretical basis

The question of how to describe the editorial decisions takes us to the root cause of our problems: the lack of a shared theoretical foundations that can be used to describe all the steps of the editorial process and can be used by all the computer tools involved in it.

Let us state clearly that the described issues are not due to fact that the implementations of the tools are incomplete.

The root cause lies, instead, in the fragile theoretical foundations upon which these tools are built. For example, it would not be too hard for a tool to automatically typeset a whole edition, but it cannot do it because it does not know which variants should be considered correct, which are relevant enough to be included in the critical apparatus and which should just be omitted. In turn, a tool cannot know how to identify a variant as correct, relevant or irrelevant if the editor has not explained it. And under the theoretical frameworks used by current tools, the editor has no way to explain to the tool what rules it should apply to identify a variant as relevant.

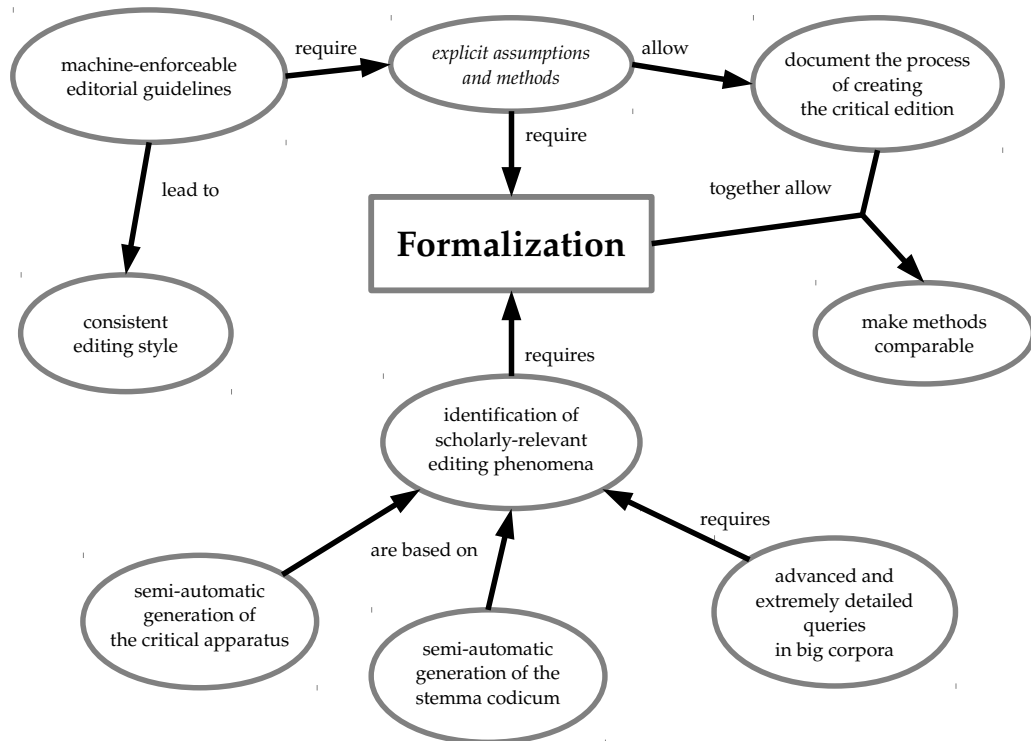


Figure 2: All the editorial steps need, directly or indirectly, a shared formalization.

It is clear that to address our problems there must be a way for the editors to describe their rules, decisions and preferences to the machine. In other words, a formalization of the editorial process. In order to be useful, such a formalization should define, in precise and machine-readable terms, the many different aspects that comprise the editorial workflow.

The shared formalization should allow the editors to, at least:

- define the basic elements that a tool can operate upon;
 - e.g., does the tool operate on letters? (and what is a ‘letter’ in its parlance? A Unicode codepoint? A grapheme?) or hieroglyphs? words? XML nodes? sequences?
- provide a way to group, classify and identify these basic elements;
 - e.g., which words are adjectives, which are names of people?
- name and define the known editing phenomena and the rules to detect them;
 - e.g., a *saut du même au même* is detected in document B, if document B contains the sequence W1 W5, while document A contains the sequence W1 W2 W3 W4 W5 and W1 is identical to W4;
- define which classes of editing phenomena are relevant and which are not;
 - e.g., orthographic variations in general = NON RELEVANT, orthographic variance in the names of kings = RELEVANT;
- state rules on how certain classes of editing phenomena influence the critical edition;
 - e.g., if document A contains a sentence similar to another found in B, but the sentence in A has been truncated due to *saut du même au même* \Rightarrow then A cannot be an ancestor of B in the stemma codicum;
 - e.g., all orthographic variants of the names of kings must appear in the critical apparatus and in bold.

Such a shared formalization is needed because all the editorial steps are based, directly or indirectly, on it. This dependence is graphically exemplified in Figure 2.

Proposed solution: structured changes, machine-readable editorial knowledge

One such shared formalization could be created borrowing existing models from computer science, in particular from the field of document changes: the Universal Delta Model (UniDM) (Barabucci 2013), the associated document model CMV+P (Content Model Variants + Physical carrier) (Barabucci forthcoming) and the concepts of meaningful changes and detection rules (Barabucci 2016).

The CMV+P document model sees digital documents as stacks of abstraction levels, each storing content according to a certain model. For example, an XML-TEI document is seen at the same time as a graph of TEI objects, as a tree of XML nodes, as a string of Unicode codepoints, as a series of UTF-8 byte sequences, as a series of bits, and so on. The content inside each abstraction level is addressed according to an addressing scheme that suits the model of that level (e.g., XPath or XPointer for XML, 0-based indexes for bytes). This precise system allows tools working at different levels of abstraction to work on the the same document without loss of precision or lossy data/model conversions.

On top of this data model, the Universal Delta Model provides a uniform way to describe the differences (the *delta*) between two or more documents. These differences (termed *changes*) can be basic or structured. Basic changes describe the simplest kind of changes that can happen at a certain abstraction level. For example, at the alphabetic level, letters can be removed or added, while at the XML level what is removed or added are elements, comments, text nodes and so on.

Basic changes can be grouped together to create structured changes if they match a certain detection rule. For example, if one *deletion* and one *addition* operate on the same node, we can construct one *replacement* structured change by combining these two basic changes. Similarly, if we see that the letter being added is the uppercase version of the letter being deleted, then we further classify this *replacement* change as a *capitalization* change.

Using this technique editors could define their own detection rules and use these rules to explain to the machine how to classify variants and what to do with the classified variants. One example of sophisticated detection rules are the rules for the detection of undo operations in the writings of Samuel Beckett (Barabucci 2016).

It is envisioned that the community of digital scholarly editors could share their rules in public repositories, letting other editors reuse their rules or write even more refined rules on top of them.

Detection rules could as well be published as part of the respective edition to make it verifiable.

Conclusions

Currently only few steps of the editorial workflow of a digital scholarly edition are automated or receive help from computer tools. The main issue with the current tools is that they do not share data models and formats: each tool uses its own idiosyncratic data model. For this reason, making the tools work together is extremely time-consuming, cumbersome and prone to information losses. This problem is also the root of many limitations: one example is the lack of feedback or communication between the tools and the editors, another example is the impossibility for editors to suggest their preferences to these tools and influence their behaviour.

We identify the root cause of this issues with the lack of a shared formalization and propose a shared formalization that is based on models and techniques borrowed from computer science, in particular from the field of document changes. Using the Universal Delta Model, the CMV+P document model and the concepts of structured changes and detection rules it is possible to define in precise and rigorous terms all the editorial decisions taken during the creation of a digital scholarly edition.

This shared formalization would lead to the semi-automatization of the editorial process, cleanly dividing the responsibilities between editors and computers. The responsibility of editors would be to describe their choices and decisions, including rules and exceptions. Computers would, instead, deal with applying these rules and decisions in the best way.

This kind of semi-automatization would leave the editors in charge of all the scholarly decisions, while handing over to the machine the more mechanical part of the work, such as normalizing the transcriptions, collating the documents, removing irrelevant variants, typesetting the edition and so on.

Additional features that this paradigm would bring are the possibility 1) to perform advanced pattern-based search; 2) to replay the past work if, for example the set of witnesses has changed, an editorial rule has been revised or a transcription has been improved; 3) to verify if the stated editorial rules have been properly followed and the end results are replicable.

References

- Andrews, Tara. 2014. 'Analysis of variation significance in artificial traditions using Stemmaweb.' *Digital Scholarship in the Humanities* 31 (3): 523-539. DOI: 10.1093/llc/fqu072.
- Barabucci, Gioele. 2013. 'Introduction to the universal delta model.' In *ACM Symposium on Document Engineering 2013, DocEng '13*, Florence, Italy, September 10-13, 2013, edited by Simone Marinai and Kim Marriott. ACM, 47-56. DOI: 10.1145/2494266.2494284.
- . 2016. 'Manuscript annotations as deltas: first steps.' In *DChanges '16 Proceedings of the 4th International Workshop on Document Changes: Modeling, Detection, Storage and Visualization*, edited by Gioele Barabucci, Uwe M. Borghoff, Angelo Di Iorio, Sonja Schimmler, Ethan V. Munson. ACM. DOI: 10.1145/2993585.2993591.
- . Forthcoming. 'The CMV+P document model.' In *Versioning Cultural Objects*, edited by Roman Bleier and Vinayak Das Gupta. (accepted for publication)
- Burghart, Marjorie. 2016. 'The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features.' *Journal of the Text Encoding Initiative* 10. DOI: 10.4000/jtei.1520.
- Capitularia*. Edition der fränkischen Herrschererlasse. <http://capitularia.uni-koeln.de/>. Accessed online 2017-04-06.
- Dekker, Ronald, Dirk van Hulle, Gregor Middell, Vincent Neyt and Joris van Zundert. 2015. 'Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project.' *Lit Linguist Computing* 30 (3): 452-470. DOI: 10.1093/llc/fqu007.
- Rosselli Del Turco, Roberto, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti, and Jacopo Pugliese. 2014. 'Edition visualization technology: A simple tool to visualize tei-based digital editions.' *Journal of the Text Encoding Initiative* 8. DOI: 10.4000/jtei.1077.
- Salus, Peter H. 1994. *A Quarter Century of UNIX*. Addison-Wesley Professional. ISBN 978-0201547771.
- Van Zundert, Joris. 2016. 'Barely Beyond the Book?' In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo. Cambridge: Open Book Publishers, 83-106. DOI: 10.11647/OBP.0095.05.

Modelling process and the process of modelling: the genesis of a modern literary text

*Elli Bleeker*¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

'She showed him the stories in statu nascendi. He could see how she filled (them) with the truffles of a mocking thought(;) how she ornamented them with the rigging pendants of metaphores ...'

(Raymond Brulez, 'Sheherazade of Literatuur als Losprijs'; own translation)

'Follow the question mark, not the exclamation point.'

(David Bowie)

Scholarly editors are a peculiar species. They know a literary work almost better than its creator does; yet they seem content with a role 'behind the scenes'. At the same time, they hope that their edition reaches a large public and succeeds in convincing its readers of the value of the text. What this 'text' may be and how it can best be represented varies according to a number of factors. One of the most important factors are the editors themselves, namely from which perspective and with what intention they approach the text. Peter Shillingsburg refers to this as the editor's 'orientation' towards the text (Shillingsburg 2001: 5).

This is an interesting given, for it describes the presence of a subjective part in a research methodology that in general includes a meticulous and careful examination of facts. A large body of literature exists on this tension between objective editing and interpretation (e.g., Zeller 1995; Gabler 2010; Pierazzo 2015). In recent years, it has become clear that objective editing is both infeasible and undesirable. The

¹ elli.bleeker@huygens.knaw.nl.

present-day literature seems to consider this a positive development: according to Paul Eggert, it relieves textual scholars from the ‘guilt’ of critically intervening between reader and text (Eggert 2005: 94). And Shillingsburg, who appears to have a slightly different idea of the character of editors, writes that at least ‘it should keep them from acting as though their work was either universally adequate or faultless’ (Shillingsburg: 3). To let go of this ‘untenable pretense of objectivity’, as David Greetham describes it (qtd. in Eggert, 84), means that editors can ‘acknowledge a more self-reflective hermeneutics’ (Dahlström 2006: 362). These opinions form, together with the technological developments and increasingly interdisciplinary editing projects, a fertile ground for more experimental approaches to editing.

The conditions outlined above also accord well with the idea behind modelling. In her extensive discussion of this concept, Elena Pierazzo stresses two characteristics: that modelling is an ‘iterative, learning process’; and – accordingly – that failure is completely acceptable (2015: 39; idem 63). Not incidentally, modelling has become an important aspect of digital editing. Its trial-and-error character indeed could result in a more self-reflective textual scholarship; a discipline that is interested in the *process* of editing and acknowledges that – just as the ‘final’ text does not exist – the ‘final edition’ of a work might be an outdated notion. This is epitomized by Edward Vanhoutte’s response to the critical reviews of his digital edition of Stijn Streuvels’ *De teleurgang van de Waterhoek*: his analysis of the reasons behind some of the edition’s shortcomings are almost as a valuable research output as the edition itself (Vanhoutte 2006: 162-5).

The current research project can be seen in a similar light. Its goal is to examine the different possibilities of digitally representing the genesis of a literary work, and what that implies for the tasks of the editor. It uses as case study the story collection *Sheherazade of Literatuur als Losprijs* (1932) by Flemish author Raymond Brulez. A study of the *_Sheherazade_-material* demonstrates what scholarly editors already know: the genesis and transmission of a literary work can be as compelling as the work itself. The extant documentary sources offer a number of promising possibilities to illustrate the development of this text, each possibly presenting its own challenges. The section that follows demonstrates one of these challenges using an example from the genetic dossier² of *Sheherazade*. Our example regards the development of the so-called ‘gramophone sentence’ and consists of the following fragments, all located in the Archive and Museum for Flemish Cultural Life in Antwerp (AMVC):

- N1: note 5/33 (B 917 / H / 2a)
- N2: note 15/33 (B 917 / H / 2a)
- MS1: front page of the draft manuscript (B 917 / H / 2b)
- MS8: page 8 of the draft manuscript (B 917 / H / 2b)
- TSfol: folio typescript with author’s corrections (B 917 / H / 3)
- TSq: quarto typescript with author’s corrections (B 917 / H / 3)
- P: page proof (B 917 / H / 3)

2 The genetic dossier, also referred to as *avant-texte*, is a selection of all extant documentary material of the work (e.g., notes, draft manuscript, corrected typescripts) critically assembled by the editor, that together gives a representation of the text’s genesis.

Note 5/33 (N5, depicted below in Figure 1), mentions the ‘flagellatory music’ of the opera by Richard Strauss.

A second, more substantial note also mentions a gramophone. On note 15/33 (N15; Figure 2), Brulez describes in his typical lyrical terms how the needle of a gramophone player ‘floats over the ebonite black whirl of a vinyl record’; the reflection of the light on the record resembling the hourglass of eternity, symbolizing the nearing end of its existence.

The fact that the gramophone plays a significant role in the story is confirmed further by the title page of the manuscript (M1, Figure 4). On this document Brulez wrote the title of the main narrative, the date and place of inspiration (August 16th, in the ‘Kursaal’ in Oostende), the period of writing (August 1929 – March 1930) and, in the lower left corner, one comment: ‘the gramophone: Pandora’s box from which Shiriar’s (dreams?) were constantly born’. This reference to Pandora’s box, however, does not recur elsewhere in the main narrative. It is also not clear when he wrote this comment, but considering the place and the slightly thicker ink it was probably in a later stage than the title and his name.

Shiriar
de grammofoonmuziek:
Strauss "Salome": muzikaal
flagellantisme
flagellantische muziek.

Figure 1: N5 with keywords ‘de grammofoon(sic) muziek: / Strauss ‘Salome’: muzikaal / flagellantisme / flagellantische muziek’.

- 22 -
Grammofoon.
De grammofoon die vlotter deinceert op den swartsten
hroei kolk der ebonieten plaat waar ^{waarschijnlijk} het licht
symboolische gelykenis met een naar een suet einde
stijgend spieglend ~~bevat~~ bestaan te volkelgen - het lichtspel
der schijn van den samelater der eeuwigheid spiegelde.

Figure 2: N15, an extended version of the gramophone sentence.

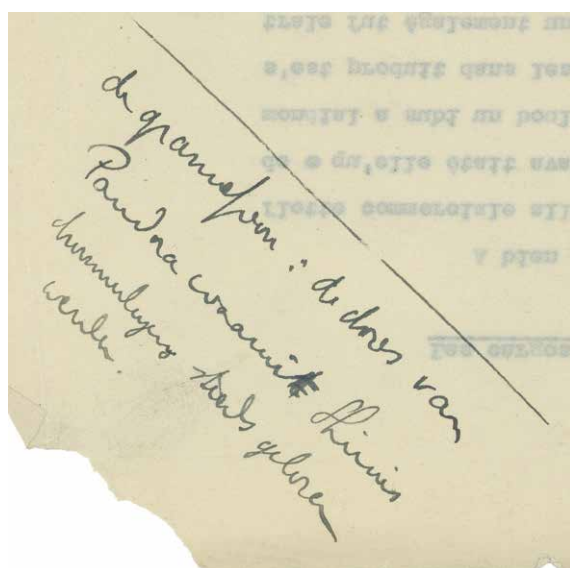
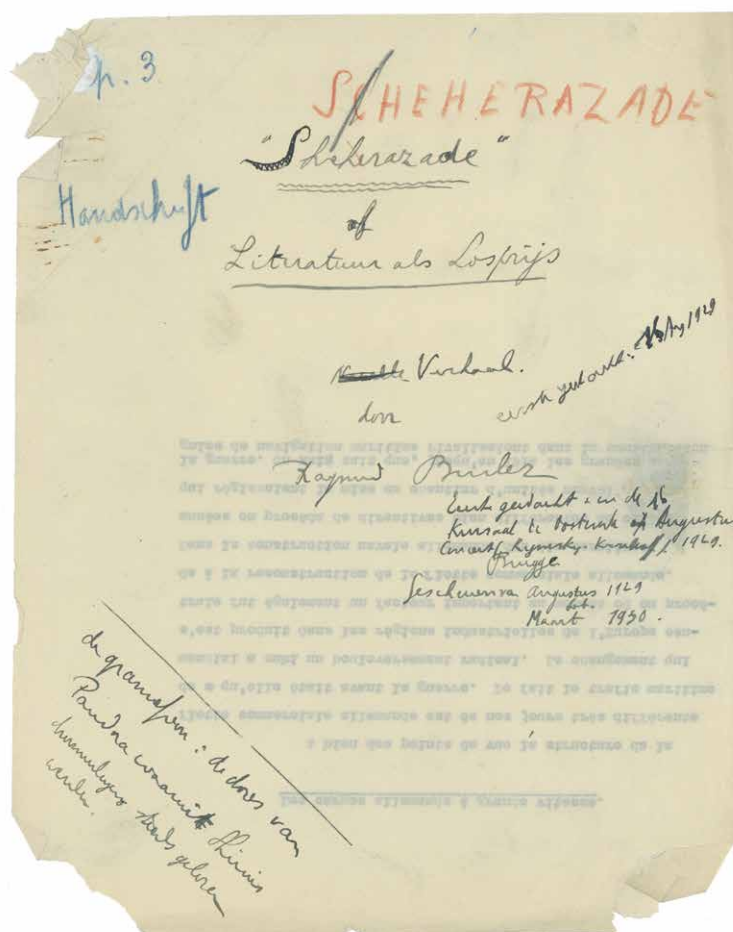


Figure 3-4: M1, the front page of the draft manuscript. In the lower right corner (in close-up, left) a reference to the gramophone.

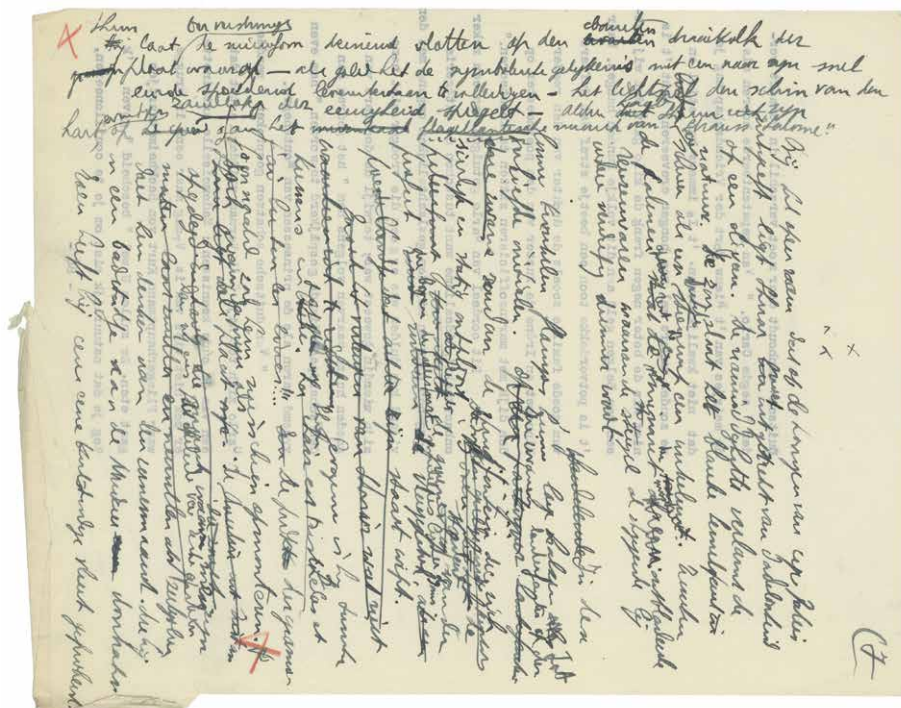


Figure 5: M8 (flipped sideways). The gramophone sentence is written in the left margin, which suggests that it was added in a later stage.

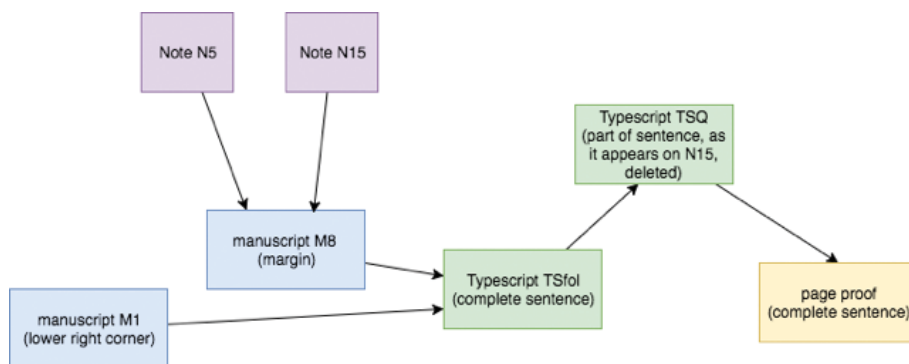


Figure 6: first conceptual model of the relationships between segments of the text.

There is no direct reference to Pandora's box in the main narrative, but the phrases on the notes are both incorporated – in slightly altered versions – into the margin of the manuscript M8 (Figure 5).

In general, Brulez' composition process started with the sketching of a writing plan, making notes, and drafting a manuscript (*cf.* Fierens 2015, 58-9). However, the fact that the gramophone sentence is added in the margin of MS8 suggests that at least N5 and N15 are written *after* the first manuscript draft. The gramophone sentence subsequently is incorporated – with slight variations – in the folio typescript (TSfol) and the quarto typescript (TSq), as well as in the page proof (P) of the story. A first conceptual model of the sentence is depicted below (Figure 6).

Despite its simplicity, this model clarifies the dynamic workings of the text and illustrates how the extant documents are related. Keep in mind that this represents only one sentence. The links between the nodes are meaningful in the sense that they represent the path between two writing stages. The model shows, among other things, the connection between the manuscript MS8 and the two notes N5 and N15 without necessarily indicating a chronological order. Textual genesis is often not linear or chronological; it can be imagined as a dynamic network of interrelated documents.

In her discussion of a model of textual transmission, Pierazzo mentions 'channels' through which text is transmitted and, during the transmission, is susceptible for 'noise' and other interferences that lead to variance (2015: 69). Although Pierazzo talks about textual transmission on a larger scale, it is possible to extend this metaphor to our case: the path between different writing stages would be the channel. The next challenge is to make sense of those relationships, whether we call them paths or channels, while retaining part of their dynamicity. The rudimentary nature of N5 adds another level of complexity: can we consider it a proper version of the gramophone sentence? How do we handle such loose snippets and similar paralipomena?

Previous studies (*cf.* Van Hulle 2008) build upon the aforementioned 'network'-metaphor and suggest recreating a similar hypertextual structure. Pierazzo suggests that 'this can be easily done in the XML (*sic*) source code, but not yet in the output'.³ So what could the output look like? If we consider the practice of a number of existing editions dealing with similar challenges, it becomes clear that there is not one 'correct' approach. There is the 'linkemic approach' of the aforementioned Streuvels-edition of Vanhoutte and Marcel DeSmedt that indeed connects related paragraphs within an Xpointer structure. Moreover, John Bryant's fluid text edition of Melville's *Typee* proposes to explain textual variation with editorial narratives. A third example is the Beckett Digital Manuscript Project that connects variant sentences by means of the same TEI XML element <xml:id>. Users of the Beckett Archive can select sentences to collate and subsequently see their connections in the collation output, where the versions are presented in a

3 In a blogpost on May 11, 2012 <<http://epierazzo.blogspot.co.uk/2012/05/genetic-encoding-at-work.html>>, last accessed on January 15, 2016.

static ‘synoptic sentence viewer’.⁴ In general, it is important to keep in mind that the goal is to provide a suitable environment to *study* the work’s genesis. This does not imply necessarily that the model needs to *mimic* it.

Although it is only one aspect of Brulez’ writing process, the example of the gramophone sentence illustrates the intriguing issue of modelling and representing textual relations across documents. Future research entails an examination of the existing editorial practices, and testing different ways to represent the fascinating stories told by textual variation. Is it possible to maintain, as Alan Galey describes it, ‘the double-vision that characterizes textual scholarship: to see at once both the signifying surface and what lies beneath’ (2010: 106)? It seems necessary, now more than ever, that scholarly editors find a balance between establishing editorial standards and respecting the ‘unique features of every single writing process’ (Van Hulle 2004: 6). As a result, they might create increasingly distinctive editions that reflect not only the writer and the work in question, but also the unique features of their editor(s).

References

- Brulez, Raymond. 1932. *Sheherazade of Literatuur als Losprijis*. Antwerpen: De Nederlandsche Boekhandel.
- Dahlström, Mats. 2006. Review of ‘Readings. Types of Editions and Target Groups.’ in *Variants* 5, edited by Luigi Giuliani, Herman Brinkman, Geert Lernout and Marita Mathijsen, 359-366. Amsterdam: Rodopi.
- Eggert, Paul. 2005. ‘These Post-Philological Days’. In *Ecdotica* 2, 80-98. Roma: Carocci.
- Fierens, Sarah. 2015. ‘Samenstelling van een tekstgenetisch dossier. De kladversies van Raymond Brulez’ Sheherazade’. MA diss., University of Antwerp.
- Gabler, Hans-Walter. 2010. ‘Theorizing the Digital Scholarly Edition.’ in *Literature Compass* 7, 43-56. Oxford: Blackwell
- Galey, Alan. 2010. ‘The Human Presence in Digital Artefacts’. In *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, edited by Willard McCarthy. Cambridge: Open Book Publishers.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing. Theories, Models and Methods*. Surrey: Ashgate.
- Shillingsburg, Peter. 2001. ‘Orientations to Text.’ In *Editio* 15, edited by Bodo Plachta and Winfried Woesler, 1-16. Tübingen: Max Niemeyer Verlag.
- Van Hulle, Dirk. 2004. *Textual Awareness: a Genetic Study of Late Manuscripts by Joyce, Proust, and Mann*. Ann Arbor: University of Michigan Press.
- . 2008. ‘Hypertext and Avant-Texte in Twentieth-Century and Contemporary Literature.’ In *A Companion to Digital Literary Studies*, edited by Susan Schreibman and Ray Siemens. Oxford: Blackwell.

4 Cf. the BDMP documentation on <http://www.beckettarchive.org/manual.jsp#conventions_tools>, last accessed on January 15, 2016.

- Vanhoutte, Edward. 2006. 'Traditional editorial standards and the digital edition.'
In *Learned Love. Proceedings of the Emblem Project Utrecht Conference (November 2006)*, 157-174. The Hague: DANS symposium Publications 1.
- Zeller, Hans. 1995. 'Record and interpretation: analysis and documentation as goal and method of editing.' In *Contemporary German Editorial Theory*, edited by Hans Walter Gabler, George Bornstein, and Gillian Borland Pierce, 17-58. Ann Arbor: University of Michigan Press.

Towards open, multi-source, and multi-authors digital scholarly editions

The Ampère platform

Christine Blondel¹ & Marco Segala²

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

André-Marie Ampère (1775-1836) is renowned for his fundamental contributions to classical electromagnetism after his mathematical theory of electrodynamics as published in 1826 (Ampère 1826). James Clerk Maxwell referred to Ampère as 'the Newton of electricity': 'The experimental investigation by which Ampère established the laws of the mechanical action between electric currents is one of the most brilliant achievements in science. The whole, theory and experiment, seems as if it had leaped, full grown and full armed, from the brain of the Newton of electricity. It is perfect in form, and unassailable in accuracy, and it is summed up in a formula from which all the phenomena may be deduced, and which must always remain the cardinal formula of electro-dynamics' (Maxwell 1973: part IV, chapter III, § 528, p. 162). But his achievements went well beyond physics: before 1820 his reputation – and his position within the prestigious Académie des Science of Paris – was mainly due to his activity in mathematics (Ampère 1802, 1815) and chemistry (Ampère 1814). In fact, Ampère was a polymath and his writings reflect various interests and research projects in natural sciences and humanities.

1 christine.blondel2@cnrs.fr.

2 marco.segala@univaq.it.

Today we still conserve the greatest part of Ampère's writings: quantitatively and qualitatively his corpus is one of the most significant in the history of French science. In 2005 the project for a digital edition of his corpus was launched: Christine Blondel, researcher at the Centre national de la recherche scientifique (CNRS), obtained a grant from the French Agence nationale pour la recherche (ANR) and the support of the Cité des sciences et de l'industrie and the Fondation EDF (the foundation of the Electricité de France, Inc.). Hence the platform 'Ampère et l'histoire de l'électricité' (www.ampere.cnrs.it) was put in place.

In 2016, thanks to another ANR grant, a new project led by Christine Blondel and Marco Segala (University of L'Aquila, Italy) at the Centre Alexandre Koyré has achieved a completely new version of the platform. The Atelier des Humanités Numériques of the Ecole Normale Supérieure at Lyon has given technical collaboration in establishing the new platform through BaseX³ and Synopsis.⁴

The first platform was projected to host two different sets of sources and documents: the first one presenting André-Marie Ampère's writings (publications, correspondence, and manuscripts), to give scholars an easy access to the author who at the beginning of the 1820s established a seminal mathematical theory of electrodynamics; the second one – mainly addressed to laypeople, school professors in physics, and their pupils – devoted to the history of electricity and magnetism and providing primary and secondary sources and multimedia documents with videos of historical experiments.

While the second part has migrated into the new platform with minor changes in content, the first part has seen substantial transformation and becomes an actual digital scholarly edition of Ampère's writings and correspondence. In the first website the publications, with the exception of a few ones, were displayed in PDF format; the correspondence mainly relied upon the De Launay's uncomplete edition (De Launay 1936: vol. 3); the manuscripts section collected the 53417 facsimiles of the papers conserved at the Archives of the Paris Académie des Sciences – but only a very small portion had been transcribed. The new version is offering the complete TEI transcription of Ampère's publications and correspondence – now completed notably by an extended youth correspondence between Ampère and another young bourgeois passionate about science – and TEI transcriptions of a qualitatively important selection of the manuscripts (in the new platform more correctly called *archives*, the French term for 'archival material'). The corpus is enriched by an indexation of all the transcribed material. Moreover the annotation software 'Pundit', developed by the Italian Net7, has been implemented to allow private and public annotations.

This presentation will illustrate contents, aims, and challenges of the new digital edition of Ampère's corpus and will consider the importance of implementing the digital edition with annotation software.

3 Open source software that is defined as «a light-weight, high-performance and scalable XML Database engine and XPath/XQuery 3.1 processor»; it «includes full support for the W3C Update and Full Text extensions». It is an interactive and user-friendly GUI frontend that gives full insight into XML documents.

4 Open source software, it is a full XML corpus publishing system from BaseX XML database.

Contents: the corpus and its TEI transcription

Ampère's corpus is composed of 153 publications (3689 pages), mainly in the domain of mathematics, physics, and philosophy of science, 1182 letters (from and to Ampère), and 53417 archival pages devoted to scientific disciplines (mathematics, physics, chemistry, astronomy, natural history), the human sciences (psychology, philosophy, linguistics), and personal writings (autobiography, personal journals, poetry).

As previously said, all the publications and letters, together with a significant selection of the manuscripts, have been encoded according to the TEI. Choosing TEI for the encoded transcriptions was not only motivated by the exigency to adhere to consolidated international standards in digital scholarly editions. The crucial point, here, is that Ampère's corpus consists of three kinds of material (publications, manuscripts, and correspondence) in different forms: publications, reprints, printed proofs; archival material with texts, drawings, calculations, tables, lists, classifications; printed letters with or without the original manuscript, unpublished manuscripts letters, anonymous or undated letters. And the complexity grows when one considers that, Ampère being a polymath, each document can be related to different topics.

It is evident that in this case the mere transcription and the related possibility of textual search is necessary but not sufficient. What a corpus like Ampère's needs is both the establishing of connections among the three kinds of documents and support to interrelated research – as required by scholarship. Opting for TEI-based transcriptions intended to add easily manageable and quickly exploitable information to the texts. It also gives the possibility of interoperability with other digital corpus of savants or philosophers of the time.

Even if the idea is sound, in Ampère's case its realization is complicated by the large quantity of documents and pages to transcribe. Maybe this is not a 'big data' project, but certainly quantity is a relevant factor and impacts every choice. When at the beginning we decided for 'light' encoding, the ultimate reason was our intention to transcribe and encode many thousands of pages (by TEI) and mathematical formulas (with MathML, Mathematical Markup Language). As we did not want to lose neither physical information (chapters, pages, paragraphs, and lines) nor editorial details (deletions, gaps, old spelling, language), we found ourselves obliged to limit semantic encoding to only five categories: people, institutions, publications titles, places, and dates. Notwithstanding such 'lightness', the transcriptions ended up with 164 tags – and some of them were not light at all: 15000 mathematical formulas; 15000 occurrences of sic/corr; 11700 <persName>; 11800 <date>; 6600 <place>; 1500 <institution>.

As a result, Ampère's writings are now collected in hundreds of XML files ready for web edition and completely searchable – even within mathematical formulas. Even if Ampère is famous as a physicist – and studies on Ampère are mainly in the domain of the history of physics – his contributions to other fields of knowledge were not negligible and we hope our transcriptions can stimulate specific attention by the historians of mathematics, philosophy, life sciences, education, and linguistics.

Aims and challenges: indexing and annotating

The present encoded transcriptions establish a digital edition of Ampère's texts that replicate the same standards through the three kinds of documents (publications, manuscripts, and correspondence). This is the first step to the exploitation of Ampère's writings: making available new materials in order to both understand and exhibit not only how he contributed to the growth of knowledge, but also to enlighten the life of an 18th-century young provincial bourgeois who became a 19th-century professor, education inspector, and academician.⁵ Main goal of this multi-source edition is to provide scholars with a homogeneous corpus that is as easy as possible to explore.

As major tools for such exploration, indexes, a faceted search engine, and an annotation software are embedded in the Ampère platform.

Indexes

Indexes are essential for exploring and analysing the entire corpus. Each index entry must link to the different passages in the three kinds of documents where the indexed entity is mentioned.

Quantitatively and qualitatively, the general name index is the most important. It is generated by the <persname>tag in the encoded transcription. Once again, the difficulties come from the dimensions of the corpus: it mentions about 2000 people in about 11700 occurrences and with different graphical forms.

Our first work consisted of relating those 11700 occurrences and their graphical varieties to identified people; later we made efforts to enrich those identified names with first names, dates, and short biographical data. We decided to rely on authority files provided by the Bibliothèque Nationale de France (DataBNF) and Wikipedia to both gain in standardization and provide interoperability with other *corpora*. It is pleonastic to say that identifying people from the occurrences required an enormous effort, but a worthy one: the merit of indexing via authority files is that once one has identified a person, the reference is stable. And this is certainly highly desirable when compared with the inherently provisional indexes of the past.

Together with the name index, two other indexes are generated by the encoding of places and institutions. Place names are of particular importance as they are provided mainly by Ampère's correspondence and travels as general inspector of the ministry of education. As such, they make available a 'geography' of Ampère's network, life, and career.

Last but not least, there is a list of the books he refers to in his correspondence – often implicitly. It highlights his strong interests in religion, philosophy, psychology, and topics usually considered at the margins of science, like animal magnetism.

5 Ampère was professor at the Ecole centrale, the Ecole polytechnique, and the Collège de France. He also work as 'inspecteur général' of the Ministry of education.

Annotation

The idea underlying this project is that digital humanities must support research in humanities in an innovative way. Traditional research in the *corpora* of great scientists generally is intended to establish relationships among different texts. Scholars devote their skills to selection, analysis and interpretation of texts. Digital access to a corpus – from any location and through a search engine and indexes – certainly gives substantial advantage but today we can demand something more from a digital edition.

When scholars study a corpus, they preliminarily annotate it, establish relations among different parts of that corpus, create classification and references. Put in another way: even before the publication of their research results, scholars produce knowledge as ‘annotations’ that usually are stored in their computers. The Ampère platform includes an annotation software⁶ that gives scholars the opportunity of anchoring their annotations to the web pages they are reading. Each scholar will choose whether those annotations will be private or public, then having the possibility to share their knowledge even before publication. As a consequence, the Ampère platform is to be intended as multi-source and *multi-authors*.

As editors of the platform, we are the first annotators of the corpus. We intend to show notions and relations that can be useful for further research. It is very interesting, for example, to expose genetic relations from germinal concepts in manuscripts or letters to mature and definitive concepts in publications.

Concluding remarks

Our view is that a *digital* scholarly edition must provide something new and something more than a *printed* scholarly edition. The classical scholarly edition is definite regarding its object and definitive; a digital scholarly edition is never definitive and it is not necessarily definite regarding its object.

We think that it must foster research by sharing knowledge and becoming itself a research tool. Our wish is that sharing annotations will establish new standards and methods in the exploitation of textual and multi-source *corpora*.

6 Pundit, by Net7, Pisa, Italy. See www.thepund.it.

References

- Ampère, André-Marie. 1802. *Considérations sur la théorie mathématique du jeu*. Lyon, Paris: Frères Périsse.
- . 1814. 'Lettre de M. Ampère à M. le comte Berthollet sur la détermination des proportions dans lesquelles les corps se combinent d'après le nombre et la disposition respective des molécules dont les parties intégrantes sont composés.' *Annales de chimie* 90: 43-86.
- . 1815. 'Considérations générales sur les intégrales des équations aux différences partielles.' *Journal de l'Ecole Polytechnique* 10: 549-611.
- . 1826. *Théorie des phénomènes électrodynamiques uniquement déduite de l'expérience*. Paris: Méquignon-Marvis.
- De Launay, Louis. 1936. *Correspondance du Grand Ampère*. Paris: Gauthier-Villars.
- Maxwell, J. C. 1873. *A Treatise on Electricity and Magnetism*. Oxford: Clarendon.

Accidental editors and the crowd

Ben Brumfield¹

Club lecture given at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.²

Thanks to DiXiT for bringing me here and thank you all for coming. All right, my talk tonight is about accidental editors and the crowd. What is an accidental editor? Most of you people in this room are here because you're editors and you work with editions. So I ask you, look back, think back to when you decided to become an editor. Maybe you were a small child and you told your mother, 'When I grow up I want to be an editor.' Or maybe it was just when you applied for a fellowship at DiXiT because it sounded like a good deal.

The fact of the matter is there are many editions that are happening by people who never decided to become an editor. They never made any intentional decision to do this and I'd like to talk about this tonight:

Digital Scholarly Editions

So all this week we've been talking digital scholarly editions, tonight, however, I'd like to take you on a tour of digital editions that have no connection whatsoever to the scholarly community in this room.

Torsten Schaßan yesterday defined digital editions saying that, 'A digital edition is anything that calls itself a digital edition.' None of the projects that I'm going to talk about tonight call themselves digital editions. Many of them have never heard of digital editions. So, we're going to need another definition. We're going to need a functional definition along the lines of Patrick Sahle's, and this is the definition I'd like to use tonight. So these are 'Encoded representations of primary sources that are designed to serve a digital research need.'

¹ benwbrum@gmail.com.

² This is a transcript of the lecture and subsequent demonstration, delivered March 17, 2016, in the Stereo Wonderland, Cologne. A video recording is available on YouTube at <https://youtu.be/7X6r-j35rE1k> and <https://youtu.be/rupktpz0Xrg> respectively.

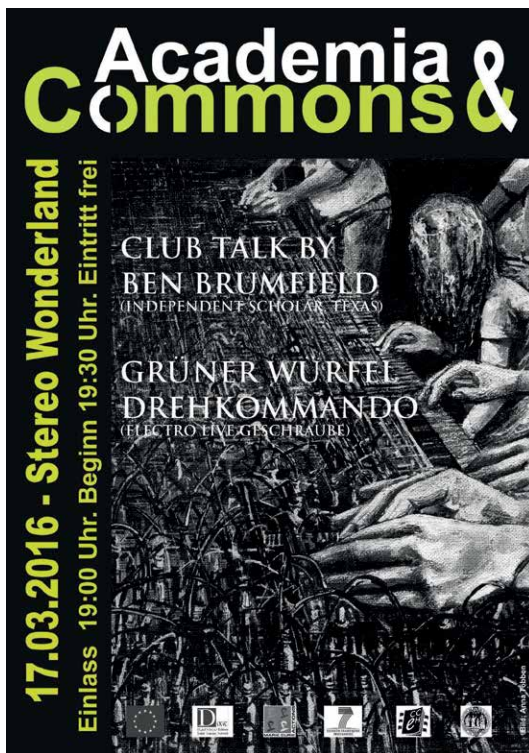


Figure 1: Club lecture and concert poster (Anna Többen / CC-BY).

All right, so the *need* is important. The need gives birth to these digital editions. So what is a *need* in the world of people who are doing editing without knowing they're doing editing? Well, I'll start with OhNoRobot. Everyone is familiar with the digital editing platform OhNoRobot, right? Right?

All right, so let's say that you read a web comic. Here's my favorite web comic, Achewood, and it has some lovely dark humor about books being 'huge money-losers' and everyone 'gets burned on those deals' (Figure 2). And now you have a problem which is that two years later a friend of yours says, 'I'm going to write a book and it's going to be great.' And you'd say, 'Oh, I remember this great comic I read about that. How am I going to find that, though?'

Well, fortunately you can go to the Achewood Search Service and you type in 'huge money-loser' and you see a bit of transcript and you click on it...

And you have the comic again (Figure 2). You've suddenly found the comic strip from 2002 that referred to books as huge money losers. Now, how is that possibly? See this button down here? This button here that says 'Improve Transcription.' If you click on that button...

You'll get at a place to edit the text and you'll get a set of instructions (Figure 3). And you'll get a format, a very specific format and encoding style for editing this web comic. All right? Where did that format—where did that encoding come from? Well, it came from the world of stage, the world of screenplays. So this reads like a script. And the thing is, it actually does work. It works pretty well. So that community has developed this encoding standard to solve this problem.

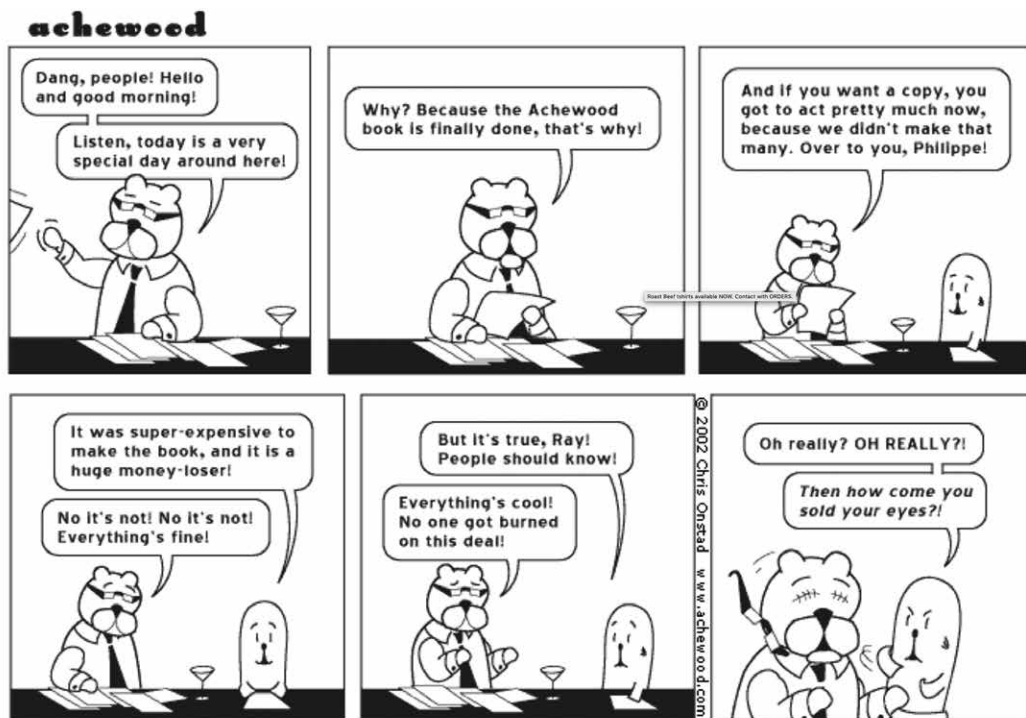


Figure 2: Achewood July 15, 2002.

Thank you for helping transcribe this comic! This helps me build a database of my comics, which makes searching for a specific comic a whole lot easier. This comic's transcription has been marked as "needing improvement", so compare it to what you find in the comic, and see if you can add or correct what's below. If you're not sure how the transcription process is meant to work, [click here](#). Thanks!

transcription for the comic at <http://achewood.com/index.php?date=07152002>:

[[A desk with papers and a martini glass. Ray, dressed in shirt and tie, is just discarding a sheet of paper.]]
 Ray: Dang, people! Hello and good morning! Listen, today is a very special day around here.

Ray: Why? Because the Achewood book is finally done, that's why!

[[Phillippe appears at Ray's side.]]
 Ray: And if you want a copy, you got to act pretty much now, because we didn't make that many. Over to you, Philippe!

Philippe: It was super-expensive to make the book, and it is a huge money-loser!
 Ray: No it's not! No it's not! Everything's fine!

Press return after each line of dialogue, and leave a blank line after each panel.
 Format dialogue like *Character's name: What are the haps?*
 You don't have to add "PANEL 1:" labels.

You can use these optional tags to add more to the transcription:
 [[Scene description]] • <<Sound effect>> • {{Meta-comic information}}
 ([click here for instructions on how to use these tags](#))

Figure 3: Achewood transcription editor.



Figure 4: Example parish register image, courtesy Free UK Genealogy CIO. Burials from 1684.

Let's say that you're a genealogist and you want to track records of burials from 1684 that are written in horrible old secretary hand (Figure 4) and you want to share them with people. No one is going to sit down and read that. They're going to interact with us through something like FreeReg. This is a search engine that I developed for Free UK Genealogy which is an Open Data genealogy non-profit in the U. K. And this is how they're going to interact with this data. But how's it actually encoded? How are these volunteers entering what is now, I'm pleased to say, 38 million records?

Well, they have rules.³ They have very strict rules. They have rules that are so strict that they are written in bold. 'You transcribe what you read errors and all!'

And if you need help here is a very simple set of encoding standards that are derived from regular expressions from the world of computer programming (Figure 5). All right? This is a very effective thing to do.

One thing I'd like to point out is that in the current database records encoded using this encoding style are never actually returned. This is (encoded) because volunteers demand the ability to represent what they see and encoding that's sufficient to do that even if the results might even be lost, in the hope that some day in the future they will be able to retrieve them.

Okay. So far I've been talking mainly about amateur editors. I'd like to talk about another set of accidental editors which are people in the natural sciences. For

3 See <https://www.freereg.org.uk/cms/information-for-transcribers/entering-data-from-registers#guidance>.

Uncertain Character Format (UCF)

Some common types of uncertainty that you are likely to encounter in your first few batches of transcription, and the technique to use for each of them, are given below. This is followed by more details of the format that we use.

Some examples

I can see one letter which could be an 'l' or a 't'.

[lt]

I can see one character which could be anything.

—

I can see two characters which could be anything.

—

I think the letter is a 'b'.

[b_]

I see a group of characters that I can't read — I don't know how many.

*

I can see two or three letters that I can't read.

_{2,3}

I can see something which could be a letter or just an ink blot.

_{0,1}

I think I see the word 'John'.

John?

Figure 5: Uncertain Character Format (UCF) from 'Entering Data From Registers', FreeREG project. <https://www.freereg.org.uk/cms/information-for-transcribers/entering-data-from-registers>.

years and years naturalists have studied collections and they've studied specimens in museums and they've gotten very, very good at digitizing things like...

This is a 'wet collection'. It's a spider in a jar and it's a picture I took at the Peabody Museum (Figure 6). In case you've ever wondered whether provenance can be a matter of horror (laughter) I will tell you that the note on this says, 'Found on bananas from Ecuador.' Be careful opening your bananas from Ecuador! Thanks to climate change and thanks to habitat loss these scientists are returning to these original field books to try to find out about the locations that these were collected from to find out what the habitats looked like 100 years ago or more. And for that these records need to be transcribed.



Figure 6: Spider in a jar, Yale Peabody Museum of Natural Science. (Photographed by author).

So here is the Smithsonian Institute Transcription Center (Figure 7). This is going to look familiar to a lot of people in the room. The encoding is something really interesting because we have this set of square notes: *vertical notation in left margin, vertical in red, slash left margin, vertical in red* all around 'Meeker'. The interesting thing about this encoding is that this was not developed by the Smithsonian. Where did they get this encoding from?

They got this encoding from a blog post by one of their volunteers. This is a blog post by Siobhan Leachman who spends a lot of time volunteering and transcribing for the Smithsonian (Figure 8). And because of her encounter with the text she was forced to develop a set of transcription encoding standards and to tell all of her friends about it, to try to proselytize, to convert all of the other volunteers to use these conventions. And the conventions are pretty complete: They talk about circled text, they talk about superscript text, they talk about geographical names. I'm fairly convinced – and having met Siobhan I believe she can do it – that given another couple of years she will have reinvented the TEI. (laughter)

So you may ask me, 'Why are we squished into the back of a room?' To make room for the swords. And we haven't talked about swords yet. So I'd like to talk about people doing what's called Historical European Martial Arts. This is sword fighting. It's HEMA for short. So you have a group of people doing martial arts in the athletic tradition as well as in the tradition of re-enactors who are trying to recreate the martial arts techniques of the past.

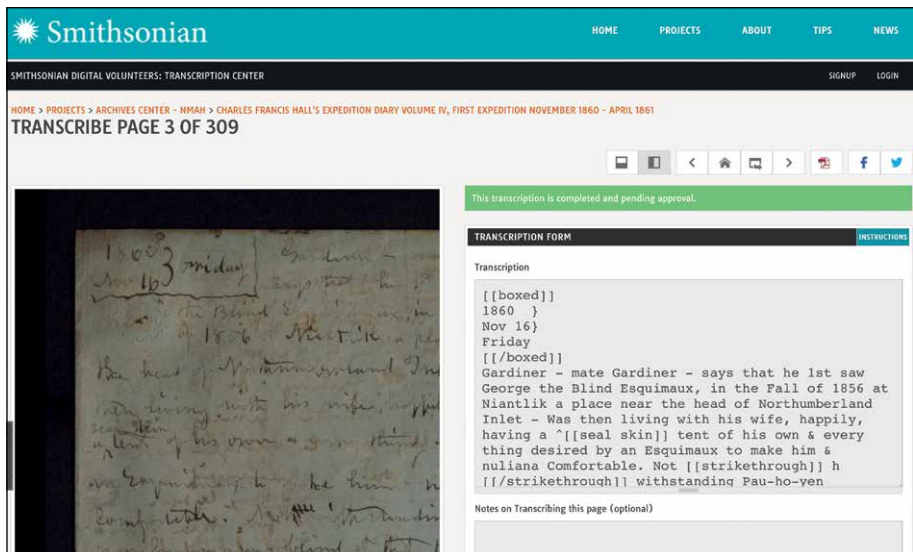


Figure 7: Smithsonian Institute Transcription Center.



Figure 8: 'Transcribing for the Smithsonian etc', blog post by Siobhan Leachman.

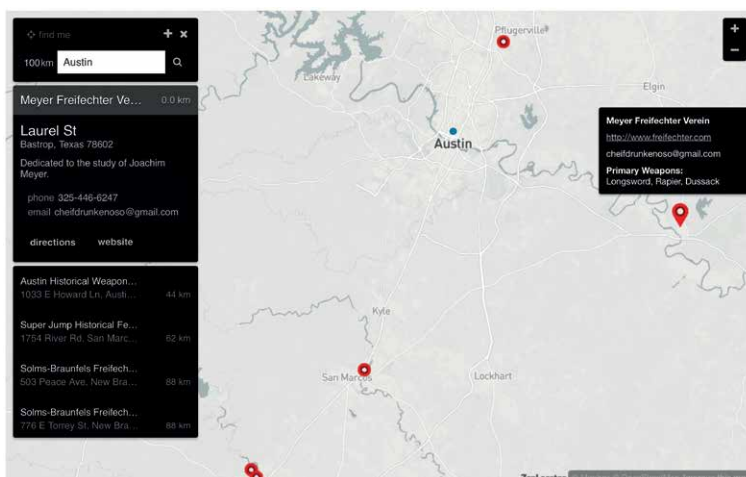


Figure 9: HEMA Alliance Club Finder.



Figure 10: Instruction (screenshot from *Back to the Source – Historical European Martial Arts* documentary by Cédric Hauteville).



Figure 11: Practicing (Screenshot from *Back to the Source*).

So there are HEMA chapters all over. This is a map of central Texas showing the groups near me within about 100 kilometers and as you can see many clubs specialize in different traditions (Figure 9). There are two clubs near me that specialize in German long sword. There's one club that specializes in the Italian traditions and there are—there's at least one club I know of that specializes in a certain set of weapons from all over Europe.

So how do they use them? Right? How do they actually recreate the sword fighting techniques? They use the texts in training. And this is a scene from an excellent documentary called 'Back to the Source', which I think is very telling, talking about how they actually interact with these (Figure 10). So here we have somebody explaining a technique, explaining how to pull a sword from someone's hand...

And now they're demonstrating it (Figure 11).

So where do they get these sources from? For a long time they worked with 19th century print editions. For a long time people, including the group in this room, worked with photocopies or PDFs on forms. Really all of this stuff was very sort of separated and disparate until about five years ago. So five or six years ago Michael Chidester who was a HEMA practitioner who was bedridden due to a leg injury had a lot of time on the computer to modify Wikisource, which is the best media wiki platform for creating digital editions, to create a site called Wiktenauer.

What can you find on Wiktenauer? Okay, here's a very simple example of a fighting manual (Figure 12). We've got the image on one side. We've got a facsimile with illustrations. We have a transcription, we have a translation in the middle. This is the most basic. This is something that people can very easily print out, use in the field in their training.




Images	Draft Translation  by Kirk Siemsen	Transcription [edit] by Bartłomiej Walczak
	A Stepping Move and an Arm Break Your opponent took a swipe with his dagger at your face. So, with an inverted right hand, continue to follow through to the dagger. During your opponent's swing, grab just under the wrist. From your left side, grasp his right elbow with your left hand, and be careful that your right hand, which stabilized his, doesn't cross over to your left side. Then, step with your left foot in front of his back foot, so you can still maintain control, and capture the dagger by bringing in your arm.	[090v] Einwerffn unnd .armpru ^{ch} . Sticht dir einer mit einnem gefau ^s stn degen oben nach dem gesicht / So far ebich au ^f mit deiner rechtn hannt ^{od(er)} ewignis deine(m) degen / Im vnder seinen stich / vorn an das glenckh / greif mit deiner lincken h[and] an seinen re: [chten] elbogen vnd wint mit der rechtn h[and] wol ybersich au ^f dein l[incke] seiten / Vnd tritt mit d[eine] l[incke] fu ^{eß} fu ^r sein bed füesß / So w[urfftst] du ⁿ in / vnd nimbst im den Degen / o[der] p[ruch] im d[en] arm /
	An Arm Break and a Stepping Move Your opponent tried to stab your face. Grab high with your inverted left hand, while securing his dagger just under the wrist. With your right hand grab from under his right elbow, and push away from you. With your left hand, bring down your opponent's hand to your left side, while at the same time step deep with your left foot behind the opponent's foot so you can maintain control as you bring in your arm.	[090v] Ein armbru ^{ch} v[on] we:rn Wiler dir dein gesicht zu ⁿ stechen / von oben so wint auf mit deiner lincken / ebichen hant / vorn unter seinen degen / an daß glennckh / Vnd greif im mit deinr re:[chten] hannt / unten an seinen rechtn elbogen / vnd ^{ebich} zeu ^{ch} an dich / vnd mit deinr l[incke] # hant dau ^{ch} im die hant au ^f dein l[incke] s[eiten] / Vnd tritt mit dein l[incke] fu ^{eß} tref hinter in ausß / so w[urfftst]u in v[nd] pri:chstu im d[en] arm /

Figure 12: Fighting manual on Wiktenauer: facsimile, translation, transcription.

Still it's a parallel-text edition. If you click through any of those you get to the editing interface (Figure 13), which has a direct connection between the facsimile and the transcript. And the transcript is done using pretty traditional MediaWiki mark up.

Page:MS Dresd.C.487 089r.png

Diese Seite enthält Bearbeitungen, die nicht zum Übersetzen freigegeben sind.

This page has been proofread, but needs to be validated.

Hie höpt sich an der ernstlich kampff zu^o roß vnd fu^oß ~

IAhie hept sich ann IMaiste~ IJohannen ILeichtenawers vechten Im harnash zu^o kampff IDaß er hât laussen schriben mitt verborgen Worten IDas stet hie nach in disem biechlin glosiert vinn vßgelegt IDas ain yede~ fechte~ vernem~en mag die kunst de~ and anderst vechten kan ~

Die vor red mitt dem text

IWer absumet /
fi vechten zu^o fu^ossen beginnet /
IDer schick sin sper /
Zway sten an heben recht

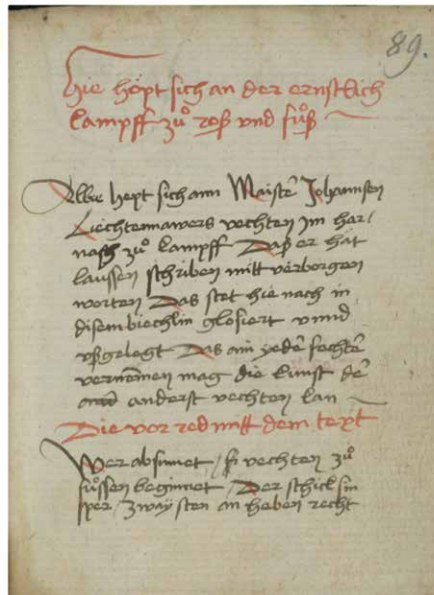


Figure 13: MS Dresd. C 487, fol. 89r (on Wiktenauer).


Images	Verse ^{is} by Mike Rasmussen Dresden Gloss ^{is} by David Rawlings	Dresden Transcription (1504-19) ^{is} by Dierk Hegedorn	Glasgow Transcription (1508) ^{is} by Dierk Hegedorn	Rostock Transcription (ca. 1570) ^{is} by Dierk Hegedorn
	In St George's name here begins the art. [1] Here begins the earnest fight on horse and foot It begins here with Mr. Johann Leichtenawer's fence in the mail coat. This he has put down in secret words. That stands now laid out and explained, therefore every fencer can understand the art, who already understands how to fence.	[H1] In sant Jorgen namen höpt an die kunst --- & fuß [H1] Hie höpt sich an der ernstlich kampff zu ^o roß vnd fuß IAhie hept sich ann IMaiste~ IJohannen ILeichtenawers vechten Im harnash zu ^o kampff IDaß er hât laussen schriben mitt verborgen Worten IDas stet hie nach in disem biechlin glosiert vinn vßgelegt IDas ain yede~ fechte~ vernem~en mag die kunst de~ and anderst vechten kan ~	[H1] Ahyn hebr sich an Dye glos vnd die auß legung der ritterlichen kunst des kampffs fechtens / Dye gedichte vnd gemacht hat Johannes Leichtenawer der ein grosser meister In der kunst gewesen ist	[H1] Hie hebrt sich an meister Johannes Leichtenawers vechtern Im harnash zu kampff das er hat laassen schreiben mitt verborgen vnd verdeckten Worten das stet hie in diesem buch glosiert vnd ausgelegt das ein lertlicher vechter vernemen mag der anders vechtern kann.
	[2] Fight with the spear He who dismounts begins fencing on foot He places his spear two stances to wield weapons right When two fight together in coats of mail, then each of them will have three different weapons: A spear, a sword and a dagger. And the beginning of the fight will occur with the spear. So you should prepare yourself with two ground positions, just as is now explained.	Die vor red mitt dem text IWer absumet / fi vechten zu ^o fußsen beginnet / IDer schick sin sper / Zway sten an heben recht [H1] wer : Gloss IMerck daß sollt du also versten Wirt zwen zu ^o fußsen in harnash mitt ein ander fechten wollen ISO soll jeder man haben dreyerlei wirt / ain sper ain schwert vnd ain degen vnd das erste anheben soll geschehen mitt dem lang langen sper Dornit sollt du dich mitt rechter wer schicken In dem anheben In zweyen stend alß du hernach höm wirst ---	Das ist der text von der vor red Wer absumet vechtern zu fuß beginnet Der schick sin sper zwey sten an heben recht wer Gloss Merck das ist / das du wissen sollt / wen zwen Im harnash zu fußsen mit ein ander vechrt sollen / So soll jlicher haben drey wer Ein sper / ein swert / vnd ein leg / vnd das erst an heben in dem kampff das soll geschehen mit dem sper / daru~ sollt du dich mitt rechter wer / gegen im mit dem sper wissen zu schicken in zwen stend /	Das ist die vorred mit dem Texte. Wer absumet, fechtens zu Fuß beginnt, der schick sein Sper, zwei sten anheben recht wer. Gloss. Merck das soltu also vernemen, wenn zween zu fuß in Harnsch mit ein ander fechten wollen, so soll Jedem haben dreyerlei wirt, ein Sper, ein schwert und ein degen und das erst anheben soll geschehen mit dem Sper, dem soltu dich mit vnsrer wirt [H1] in dem anheben zu schicken, in zweyen stend alß du hernach horen wirst.
	[3] The first ground position Spear and point then before stabs, stab without force Spring wind attack him onward disengage to face him on When you are both down from the horses, then stand with your left foot forward and hold the spear ready to throw. And close to him thus: so that the left foot always stays in front. And wait, so that you can throw before him. And follow on at once shooting forward with the sword, then he cannot safely cast against you, and grip the sword.	Der text von zweyen stend~ ISper vnd ort / den vordrich Stich on forcht ISpringe winde sech recht an / wert er tuck daß gesig im an : Gloss IDaß ist der erst stand mitt dem sper [H1] Wann ir bajow von den rossen abgetreten sind ISO stand mitt dem linken fuß vor vnd halt din sper zu dem schuß / vñ Irn also zu ^o im daß alweg din linker fuß vor blyß IVñ wart dz du ee schüßst den er vñ folg bald dem schuss nach zu ^o im mitt dem schwert ISO kan er keine gewissen schuß vñ dich haben Vñ griff zu ^o dem schwert ~	Das ist der text des ersten Stands mit dem sper Sper vnd ort / den vordrich nym an forcht / Gloss Wen du hast dein sper vnd er das sein / so schick dich mit dem erst standt gegen Irn also / Ste mit dem linken fuß vor / vnd halt dein sper in der rechte hant zu dem schuß / vnd schewe den vordrich an alle vordt / vnd folg pald dem schuß nach zu im mit dem schwert / So mag er zu dir mit dem sper keinen gewissen schuß gehab / vnd wie du mit dem schwert soll vechrt gegen dem sper / das vñdest du hernach geschribt /	Das ist der Text von den zweyen stendern. Sper vñd ort, den vordrich stich one forcht. Spring, wind sech recht an, wert er tuck das gesig im an. Gloss. Merck das ist der erst stand mit dem Sper, Wenn ir bald vñn den Rossen abgetreten seilt so steh mit dem linken fuß vor, vnd halt dein sper zu dem schuß, vñ Irn also zu im, das alweg dein linker fuß vor blyß, vñ wart das du ee schußst den er, vñd folg bald dem schuß nach zu im mit dem schwert, so kan er keinen gewissen schuß auff dich habenn, vñd greif zu dem schwert.

Figure 14: Sigmund Schining ain Ringeck edition on Wiktenauer.

Okay. Now, and I apologize to the people in the back of the room because this is a complex document (Figure 14). We get into more complex texts. So this is a text by someone named Ringeck and here we have four variants of the same text because they're producing a variorum edition. In addition to producing the variants... They have a nice introduction explaining the history of Ringeck himself and contextualizing the text. What's more, they traced the text itself and they do stemmatology to explain how these texts developed.

And in fact even come up with these these nice stemmata graphs (Figure 15).

So how are they used? So, people study the text, they encounter a new text and then they practice. As my friends last night explained to me, the practice informs their reading of the text. They are informed deeply by *die Körperlichkeit* – the actual physicality of trying out moves.

The reason that they're doing this is because they're trying to get back to the original text and the original text is not what was written down by a scribe the first time. The original text, this *Urtext*, is what was actually practiced 700 years ago and

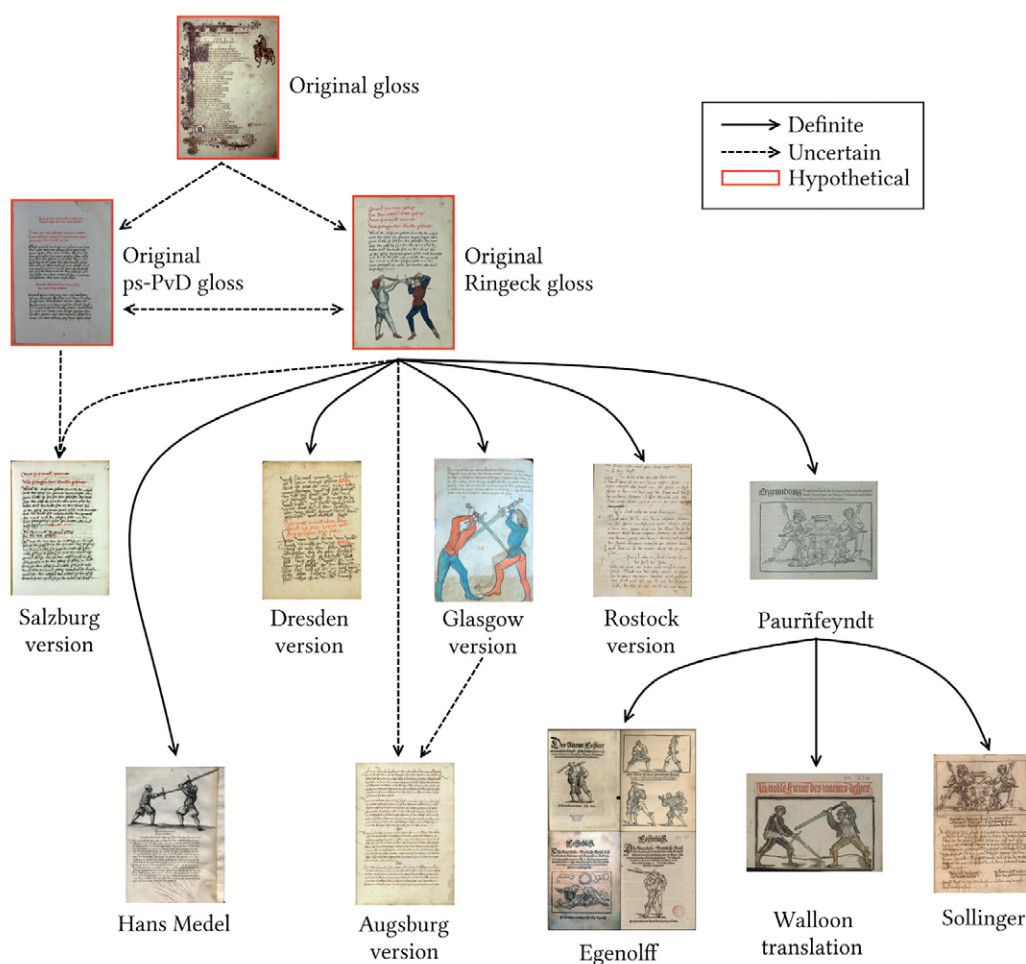


Figure 15: Provisional stemma codicum for Ringeck (Wiktenauer).

taught in schools. Much like Claire Clivaz mentioned talking about Clement of Alexandria: You have this living tradition, parts of it are written down, those parts are elaborated by members of that living tradition and now they're reconstructed.

What if your interpretation is wrong? Well, one way they find out is by fighting each other. You go to a tournament. You try out your interpretation of those moves. Someone else tries out their interpretation of those moves. If one of you would end up dead that person's interpretation is wrong. (laughter) (People think that the stakes of scholarly editing are high.)

What are the challenges to projects like *Wiktenauer*? So one of the projects -when I interviewed Michael Chidester he explained that they particularly, editors in the U.S., actually do struggle and they would love to have help from members of the scholarly community dealing with paleography, dealing with linguistic issues, and some of these fundamental issues.

One of the other big challenges that I found is – by contrast with some of the other projects we talked about – in many cases the texts on Wiktenauer are of highly varied quality. They try to adjust for this by giving each text a grade, but if an individual is going to contribute a text and they're the only one willing to do it, you sort of have to take what they get. My theory for why Wiktenauer transcripts may be of different quality from those that you see on the Smithsonian or that genealogists produce is that for those people the transcription – the act of working with the text – is an end in itself whereas for the HEMA community the texts are a way to get to the fun part, to get it to the fighting.

And now—speaking of 'the fun part' – it's time for our demonstration. It gives me great pleasure to welcome Langes Schwert Cologne, with:

Junior Instructor, Georg Schlager

Senior Instructor, Richard Strey

Head Instructor, Michael Rieck

RICHARD: Okay, two things first I will be presenting this in English even though the text is in German, but you won't really have to read the text to understand what's going on. Also, we will have to adjust for a couple of things. A sword fight usually starts at long range unless someone is jumping out from behind a bush or something. So we'll have to adjust for that. In reality moves would be quite large. All right. So he could actually go to the toilet and then kill me with just one step. So we will be symbolizing the onset of the fight by him doing a little step instead us both taking several steps. All right.

So basically, this is what it's all about. These techniques also work in castles and small hallways. (laughter) All right. Now, again we are Langes Schwert we have been doing the historical German martial arts since 1992. We train here in Cologne and today we would like to show you how we get from the written source to an actual working fighting. You can all calm down now from now on it's going to be a slow match. So, in case you didn't get what happened the whole thing in slow.

Okay, so how do we know what to do? We have books that tell us. For this presentation we will be using four primary sources: fencing books from around 1450 to 1490 all dealing with the same source material. On the right hand side you can see our second source the text you see there will be exactly what we are doing

now. Also, we use a transcription by Didien de Conier from France. He did that in 2003, but since the words do not change we can still use it. All right, so how do we know what to do? I can talk in this direction.

He's the good guy, I'm the bad guy.

GEORGE: We can see that. (laughter)

RICHARD: So how does he know what to do? I'll be basically reading this to you in English, we've translated it. In our group we have several historians. Several other members as well can actually read the original sources, but still in training we go the easy and do the transcription. But still usually we have the originals with us in PDF format so in case we figure, 'Well, maybe something's wrong there,' we can still look at it. For the most part that doesn't really matter, but the difference between *seine rechte seite* and *deine rechte seite* - 'his right side' and 'your right side' can make a difference. (laughter)

Okay, so what we're dealing with today is the easiest cut, the wrath cut. The sources tell us that the wrath cut breaks all attacks that from comeabove with the point of the sword and yet it's nothing but the poor peasant's blow. Essentially what you would do with a baseball bat, all right? But very refined. (laughter) So, usually his plan would be to come here and kill me. Sadly, I was better than him. I have the initiative so he has to react. And it says, do it like this, if you come to him in the onset, symbolized, and he strikes at you from his right side with a long edge diagonally at your head like this... then also strike at him from your right side diagonally from above onto his sword without any deflection or defense.

So that's his plan. It would be a very short fight if I didn't notice that. (laughter) So, thirdly it says if he weak—oh no, if he is soft in the bind which implies that I survive the first part which looks like this then let your point shoot in long towards his face or chest and stab at him. See we like each other so he doesn't actually stab me. (laughter)

Okay, it says this is how you set your point on him. Okay, next part, if he becomes aware of your thrust and become hard in the bind and pushes your sword aside with strength... then you should grip your sword up along his blade back down the other side again along the blade and hit him in the head. (laughter) All right, this is called taking it all from above. Okay. This is the end of our source right here. Now we have left out lots of things. There are a lot of things that are not said in the text. For example it says if he's weak in the bind, or soft in the bind, actually I'm not, I'm neutral and then I become soft. How do I know this? It doesn't say, so right here. Well maybe I could just try it being soft. It would basically look like this.

It doesn't really work. (laughter) Now, if being soft doesn't really work maybe being hard does. So I'll try that next. It doesn't really work either. Okay. So this is an example from fencing, from actually doing it you know that in the bind you have to be neutral. If you decide too early he can react to it and you can change your mind too fast.

All right, now what we read here is just one possible outcome of a fight. A fight is always a decision tree. Whenever something happens you can decide to do this or do the other thing and if you fight at the master level which is this you lots and lots and lots of actions that happen and the opponent notices it, reacts accordingly

and now if you were to carry out your plan you would die. So you have to abort your plan and do something else instead. So now we'll show you what our actual plans were and what happened or didn't happen. So, I was the lucky guy who got to go first. I have the initiative. So my plan A is always this.

And then I'll go have a beer. (laughter) And the talk is over, but he's the good guy so he notices what happened and his plan is this. He hits my sword and my head at the same time. So if I just did what they do in the movies notice that he's going bang, he is not dead, I'll do something else, no it doesn't work. I'm already finished. So I have to notice this while I'm still in the air. Abort the attack, right? I was going to go far like this, now I'm not going to do that. I'm going to shorten my attack and my stab. And from here I'm going to keep going. He is strong in the bind so I will work around his sword and hit him in the face. How to do this is described I think two pages later or three.

All right. Now, remember I was supposed to be weak or soft in the bind. My plan was to kill him, it didn't work. I had to abort it. When I hit his blade he was strong, I go around it which makes me soft which is what is says there. Okay, sadly he doesn't really give me the time to do my attack and instead, he keeps going. I was going to hit him in the face, but since I was soft he took the middle, hits me in the head. All thrusts are depending on the range. In here we do not have much range so we'll always be hitting. If we're farther apart it will be a thrust. Okay, but I notice that, so I'm not afraid for my head I'm afraid for the people. (laughter)

MAN #1: Divine intervention!

RICHARD: So I notice him hitting me in the head so I'll just take the middle and hit him in the head. It usually works. Okay. Now, obviously, the smart move for him is to just take his sword away, let me drop into the hole because I was pushing in that direction anyway and then he'll keep me from getting back inside by just going down this way. And that, basically, is how the good guy wins. Except, of course, there is a page over there where it tells me how to win. And it says, well if he tries to go up and down you just go in. See, you have to stand here.

So, basically there is never any foolproof way to win. It's always a case of initiative and feeling the right thing. Actually, there is one foolproof way, but we're not going to tell you. (laughter) This concludes our small demonstration. (applause) I'm not finished yet. So, what we did was about 60 to 70% speed. We couldn't go full speed here because of the beam. Also it was just a fraction of the possible power we could do it at. We counted yesterday, we had been for the practice session what you just saw are nine different actions that are taken within about one-and-a-half seconds and that's not studied or choreography. In each instance you feel what is happening. You feel soft, hard, left, right, whatever, it's not magic, everyone does it. Well, all martial arts do it once they get to a certain level. So we would like to thank Didier de Conier who, unknowingly, provided us with the transcriptions. We would like to thank Wiktenauer even though we do not need it that often because we can actually read this stuff. It's a great resource for everyone else and as always we have that. And oh yeah, we train at the Uni Mensa every Sunday at 2. So whoever wants to drop by and join is invited to do so. (applause)

References

- Achewood. 2002. Chris Onstad, July 15. <http://achewood.com/index.php?date=07152002>.
- Hauteville, Cédric. 2015. 'Back to the Source Historical European Martial Arts.' Documentary, YouTube. <https://www.youtube.com/watch?v=7DBmNVHTmNs>.
- FreeReg. 1998-2017. <https://www.freereg.org.uk/>.
- HEMA Alliance Club Finder. <https://www.hemaalliance.com/club-finders>.
- Langes Schwert Cologne. <http://www.langes-schwert.de/>.
- OhNoRobot / Achewood Search Service. <http://www.ohnorobot.com/index.php?comic=636>.
- Smithsonian Institute Transcription Center. <https://transcription.si.edu/>.
- Text Encoding Initiative (TEI), P5 Guidelines. <http://www.tei-c.org/Guidelines/P5/>.
- Transcribing for the Smithonian etc. <http://siobhanleachman.tumblr.com/>.
- Wiktenauer – The free library of Historical European Martial Arts books and manuscript. <http://wiktenauer.com/>.

Toward a new realism for digital textuality

*Fabio Ciotti*¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

In this theoretical paper I expose and defend a position that goes against the tide of the most common assumptions shared by the community of digital textual scholars: the time has come to go back to a *realistic* notion of (digital) textuality.

The starting point of my thesis is this: the theoretical debate about digital textuality has been deeply influenced by the post-modernist theory. If this influence was quite apparent in the debates and theories about the hypertext in the early nineties of the last century (Landow 1997; McGann 2001), we can find some of its fundamental tenets also in those fields of the Digital Humanities that could be considered less susceptible to the lures of Theory – to quote J. Culler (1997) – like text encoding and digital scholarly editing. As a consequence, even in these areas we can find a general and strong support to *anti-realist*, *interpretativist* or *social constructionist* notions about textuality and its digital representation.

Probably the most well-known formulation of this belief is the one made by Jerome McGann some fifteen years ago: 'What is text? I am not so naive as to imagine that question could ever be finally settled. Asking such a question is like asking 'How long is the coast of England?'" (McGann 2002). But we can find many sentences in the work of the most important and influential scholars in Digital Humanities domain stating and trying to demonstrate these tenets, with various kinds of arguments, ranging from empirical evidences taken from real word textual documents (especially authorial handwritten manuscripts) to theoretical speculations about the ontological nature of textuality or the epistemological conditions of our knowledge 'of' and 'about' texts.

On the contrary, the realist theory of textuality that Renear, DeRose and others had proposed under the famous formula of an *Ordered Hierarchy of Content*

¹ fabio.ciotti@uniroma2.it.

Object in the 90s (DeRose *et al.* 1990) has been harshly criticized (see for instance McGann 2001), even by its own creators that have progressively moved toward a pluralistic notion of textuality (Renear, Durand, and Mylonas 1996).

In recent years perspectivist and interpretativist theories about the nature of (digital) textuality have been reaffirmed – among which, my own (Ciotti 2001). Probably the most known pluralistic theory (and model) of textuality to date is the one proposed by Patrick Sahle (2013a) which states that the definition of text depends on how we look at it, on the aspects we are most interested in making explicit in our modeling efforts and the tacit knowledge invested in those efforts: ‘Text is what you look at. And how you look at it’ (Sahle 2013b). And from this assumption he derives a well-known wheel model of textuality based on a six dimensional space (see also Fischer 2013).

Still more recently in an article by Arianna Ciula and Cristina Marras, devoted to the meta-analysis of the modeling practices in the Digital Humanities, we can read this clear statement (Ciula and Marras 2016):

Texts are or at least have been so far the privileged objects of modeling activities in DH. Here we chose to focus on some interrelated aspects of the study of texts, which seem to us particularly relevant to exemplify their complexity and openness with respect to modeling in DH: ‘dynamicity’, ‘multidimensionality’, historicity, and processuality.

New Realism and the Text

In this paper I do not want to delve into each and every one of these theories, trying to find specific counterarguments or to deny the factual assertions related to specific complex textual artefacts. My proposal to adopt a realistic theory of digital textuality is a purely speculative argument, along the lines of the ‘new realism’ movement in the philosophical debate promoted in Italy by Maurizio Ferraris (Ferraris, Bilgrami, and De Caro 2012).

The basic idea of ‘new realism’ is that even if we do not believe (or we believe only to a certain extent) that natural sciences are the ultimate measure of truth and reality, it does not follow that we should abandon the notions of reality, truth or objectivity, as it was posited by much of the 20th century continental philosophy, and by most of the *Theory* in cultural studies and social sciences. Rather, it means that philosophy, as well as other humanities disciplines, have something important and true to say about the world and the ontological nature of their objects, and that we can find criteria to assess the truthfulness of these assertions. New realism’s fundamental tenet is that reality is given first, and only at a later time may be interpreted and, if necessary, transformed.

One of the theoretical underpinnings of new realism as conceived by Ferraris that I find useful in relation to what I want to say here, is the reaffirmation of a distinction between ontology and epistemology (a distinction that post-modernist thought has blurred if not refuted at all). The domain of ontology is the domain of what really exists, and is composed by individuals; that of epistemology is what we know about the reality, and is composed by objects (Ferraris 2016, 4):

l'ontologia è composta da individui, l'epistemologia si riferisce a oggetti (...). Questi possono essere degli individui canonici, individui in senso stretto e rigoroso, come avviene nel caso degli oggetti naturali, che esistono nello spazio e nel tempo indipendentemente dai soggetti conoscenti, e in quello degli oggetti sociali, che esistono nello spazio e nel tempo ma dipendentemente dai soggetti conoscenti – in effetti, gli oggetti sociali manifestano una dipendenza generale dai soggetti, però non una dipendenza particolare, non dipendendo da uno specifico soggetto (...). Ma gli oggetti possono anche essere individui atipici, come avviene nel caso degli oggetti ideali che, esistendo fuori dello spazio e del tempo indipendentemente dai soggetti (...). Una quarta (e ultima) famiglia di oggetti epistemologici è costituita dagli artefatti. Che sono dipendenti dai soggetti quanto alla loro produzione (...), ma che (...) possono continuare a esistere anche in assenza di soggetti.

This distinction has important consequences on the notion of truth: in fact, truth depends on the ontological state of things, as much as epistemology (our knowledge of things) has a dependence relationship with ontology (the way the world is). The state of things which is the reference of a proposition subsist independently from the proposition itself, and from any knowledge or belief that we can hold about it (Ferraris 2016, 3).

I think that we can derive a realist foundation of the notion of textuality moving from this general theoretical framework. Or to reverse McGann allegory cited before, I think that we can find a way to measure (with a certain approximation...) the length of the cost of England!

If we analyze the term 'text' in its common usage, we realize that it can be used to refer to different entities, as Sahle (and many others) has rightly observed: for example, the material document, as a synonym of 'book' (the reverse is valid, as well); the linguistic discourse fixed on a material document; the intellectual work that is constituted by that discourse. Nonetheless, out of this plurality, in ordinary language we always manage to express identity statement about texts, or to speak in meaningful way about a specific text. And we actually manage to do so because, although from an ontological point of view texts exist only as a set of individual material artefacts (the documents), when they become objects of knowledge a unitary social object comes out of this plurality. The text *emerges* from the sum of the ontological individual documents that vehiculate it, but is not less real: as a social object, it is something that exists in time and space independently from the subjects that have access to it.

Note that here I am not negating the whole pluralist view of textuality: I am only denying the unlicensed (and undesirable, in my view) consequence that texts are not really existent objects. The fact that we can describe reality at different levels does not imply that the objects we describe do not exist *in se*: this fallacy is a direct consequence of the confusion between ontology and epistemology, a confusion that I want to get rid of.

A principium individuationis for text(s)

If text is an existent, a part of reality that precedes any human reading (not to say interpretation), is there a *principium individuationis* of this text that can prove useful also in digital textual sciences (in the sense that it can be the theoretical foundation for a sensible data model for textuality)?

I think that we can find such a foundation in the suggestions of the great Italian philologist and literary theorist Cesare Segre: according to Segre a text is the ‘invariant’ emerging from every operation of material reproduction of the sequence of graphic symbols (Segre 1985, 29; translation mine):

If we consider the graphic signs (letters, punctuation etc.) as signifiers of sounds, pauses etc. . . and reflect on the fact that these signs can be transcribed many times in many ways (for instance with a different handwriting, and different characters) without changing its value, we therefore can conclude that the text is the invariant, the succession of values, compared to the variables of characters, writing etc. We also can talk about meanings, if we specify that we refer to the graphic signified, those of the series of letters and punctuation signs which constitute the text. The text is therefore the fixed succession of graphic meanings.

A text is the *invariant* in every operation of material reproduction of the sequence of graphic symbols. This definition of the text as an invariant can be connected to the analysis of the *allographic* nature of textuality provided by the philosopher Nelson Goodman (1968). This definition of text as an allographic abstract object seems to provide a criterion to identify a text on the basis of the principle of identity by substitution. The permanence of the text as invariant represents the guarantee of its identification and reproducibility. As Goodman noticed (Goodman 1968, 115):

Differences between them in style and size of script or type, in color of the ink, in kind of paper, in number and layout of pages, in condition etc., do not matter. All that matters is in what may be called ‘sameness of spelling’: exact correspondence as sequences of letters, spaces, punctuation marks.

To be more precise, the possibility to establish such identity is determined by the formal characteristics of the symbols used to make and to reproduce the text, the characters and the alphabetic writing (Goodman 1968, 116):

To verify the spelling or to spell correctly is all that is required to identify an instance of the work or to produce a new instance. In effect, the fact that a literary work is in a definite notation, consisting of certain signs or characters that are to be combined by concatenation, provides the means for distinguishing the properties constitutive of the work from all the contingent properties – that is, for fixing the required features and the limits of permissible variations in each.

Someone might notice that there is something missing in this definition. For example, the title of a paragraph is certainly a sequence of alphabetical characters: nevertheless, not even a simple, naive reader would accept that 'being title' is not important for that sequence; on the contrary, a title often plays a very important role in understanding a text, be it poetic or narrative, an essay or scientific prose. However we easily can extend the definition of text as an invariant to include these features. For example, the theory of *ordered hierarchy of content objects* (OHCO; DeRose *et al.* 1990) is an excellent example of this extension. And it is not by coincidence that among the arguments to support this theory the criterion of identity by substitution was quoted explicitly.

Another possible counter-argument against the effectiveness of the identity of spelling (and of structural composition) as an objective criterion for text identification concerns the determination of the characters on which it is based. In fact, the identification of a character on a written text implies theoretical assumptions and interpretation: the assumption that a given graphic trace, 'A', is a token of a given class of abstract traces which we identify as the 'A' character. The identification of the characters in a text is therefore a *hypercoded abduction*, as Umberto Eco would put it (Eco and Sebeok 1983, chap. 10). If, in most cases, anybody can make this abduction automatically, there are *some* cases where it is not as simple or it is impossible, and more interpretative efforts are needed: 'To identify a given phenomenon as the token of a given type, implies a few hypothesis about the expressive context and the discursive co-text' (Eco 1990, 237). In short, to recognize a character we must apply different levels of competence:

- the knowledge of the writing's notation code
- the knowledge of the verbal code
- the intervention of contextual and circumstantial competencies
- the attribution of communicative intentions to the text or the author

As a matter of fact this is exactly the sort of processes that actually happen naturally for all human beings that have the minimal set of cognitive capacities required, in almost all of the cases! And even the difficult ones are treated by the specialists by the way of a *de-automatization* of the process itself, not by some kind of magic or mysterious hermeneutic trick! Making sense of reality around us by the way of the *intentional stance* is a strategy that our species has gained and fine-tuned during evolution, and has proved to be a well-crafted mechanism in almost all the cases, as the philosopher Daniel Dennet has observed (Dennett 1990):

We rely on the norms for the formation of letters on the one hand, and on the other hand we have expectations about the likelihood that this inscription was produced with some communicative intent. We try to see a message in any string of letters (...).

So, to come to a conclusion, I think that there is no bad metaphysic in the idea that texts are really existent objects, that we can identify, refer to and extract meaning from, be they digital or not.

References

- Ciotti, Fabio. 2001. 'Text Encoding as a Theoretical Language for Text Analysis.' In *New Media and the Humanities: Research and Applications. Proceedings of the First Seminar 'Computers, Literature and Philology', Edinburgh, 7-9 September 1998*, edited by Domenico Fiormonte and Jonathan Usher. Oxford: OUCS, University of Oxford.
- Ciula, Arianna, and Cristina Marras. 2016. 'Circling around Texts and Language: Towards 'pragmatic Modelling' in Digital Humanities.' *Digital Humanities Quarterly* 10. 3. <http://www.digitalhumanities.org/dhq/vol/10/3/000258/000258.html>.
- Culler, Jonathan D. 1997. *Literary Theory a Very Short Introduction*. Oxford University Press. Oxford.
- Dennett, Daniel C. 1990. 'The Interpretation of Texts, People and Other Artefacts.' *Philosophy and Phenomenological Research* 50: 177-94. DOI: 10.2307/2108038.
- DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. 'What Is Text, Really?' *Journal of Computing in Higher Education* 1. 2: 3-26.
- Eco, Umberto. 1990. *I limiti dell'interpretazione*. Milano: Bompiani.
- Eco, Umberto, and Thomas A. Sebeok. 1983. *The Sign of Three: Dupin, Holmes, Peirce. Advances in Semiotics*. Indiana University Press. <https://books.google.it/books?id=36GaAAAAIAAJ>.
- Ferraris, Maurizio. 2016. *Emergenza*. Torino: Einaudi.
- Ferraris, Maurizio, Akeel Bilgrami, and Mario De Caro. 2012. *Bentornata realta : il nuovo realismo in discussione*. Torino: Einaudi.
- Fischer, Franz. 2013. 'All Texts Are Equal, But...' *The Journal of the European Society for Textual Scholarship* 10: 77.
- Goodman, Nelson. 1968. *Languages of Art : An Approach to a Theory of Symbols*. Indianapolis: Bobbs-Merrill.
- Landow, George P. 1997. *Hypertext 2. 0: The Convergence of Contemporary Critical Theory and Technology*. 2nd ed. Baltimore, Md: Johns Hopkins University Press.
- McGann, Jerome. 2001. *Radiant Textuality: Literature after the World Wide Web*. Palgrave Macmillan.
- . 2002. 'Dialogue and Interpretation at the Interface of Man and Machine. Reflections on Textuality and a Proposal for an Experiment in Machine Reading.' *Computers and the Humanities* 36.1: 95-107.
- Renear, Allen, David Durand, and Elli Mylonas. 1996. 'Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies.' In *Research in Humanities Computing 4: Selected Papers from the 1992 ALLC/ACH Conference*, edited by Susan Hockey and Nancy Ide, 263-80. Oxford: Oxford University Press. <http://www.stg.brown.edu/resources/stg/monographs/ohco.html>.
- Sahle, Patrick. 2013a. *Digitale Editionsformen: Textbegriffe Und Recodierung. Vol. 3. BoD – Books on Demand*.
- . 2013b. 'Modeling Transcription.' *Knowledge Organization and Data Modeling in the Humanities: An Ongoing Conversation*. August 29. <https://datasymposium.wordpress.com/sahle/>.
- Segre, Cesare. 1985. *Avviamento all'analisi del testo letterario*. Torino: G. Einaudi.

Modelling textuality: a material culture framework

Arianna Ciula¹

*Closing keynote given at 'Academia, Cultural Heritage, Society'
DiXiT Convention, Cologne, March 14-18, 2016.*

This DiXiT convention was no exception in showcasing how some of us in Digital Humanities might lean towards a preferred interest on tools rather than epistemologies (how we come to know) or vice versa. That said, as others have stated amply, the intersection between the two is where the most fruitful prospect for the field lies. This closing keynote departed from some contextual definitions – digital humanities, modelling, material culture, textuality – to reflect on and exemplify a material culture framework to modelling textually. Does it speak to the DiXiT fellows and do they have anything to say?

Modelling

In itself a polysemic word, modelling is considered to be a or *the* core research methodology in Digital Humanities (McCarty 2005). In many other research contexts, modelling is understood as a research strategy and, in particular, a process by which researchers make and manipulate external representations to make sense of objects and phenomena (Ciula and Eide 2015).² The iterative experimental cycles of modelling have been theorised extensively within industrial

¹ arianna.ciula@roehampton.ac.uk.

² Inspired by a recent article (Kralemann and Lattmann 2013), recent research co-authored with other colleagues (Ciula and Marras 2016; Ciula and Eide 2017) we argue that a semiotic understanding of models as icons and of modelling as creating/interpreting icons to reason with could be helpful to grasp the dynamic relations between models, objects and interpretations. In Ciula and Eide (2014 and 2017) we gave examples of DH modelling practices in specific contexts where we analyse the representational function of models with respect to the objects they signify as well as the inferential and epistemological processes involved in these efforts. It seems a promising avenue to investigate further.

design practices engaged with making things. The widespread use of computers in modelling (in Digital Humanities in particular) tightened up but also freed the constraints of formal modelling. What is interesting to note is that even in very technical settings – such as the one exemplified by the concerns over projects documentation in Alex Czmieł's paper – formal and informal models co-exist and interact to give sense to our modelling efforts.

Material culture

While in apparent terms an oxymoron, insightful studies in anthropology and ethnography remind us that 'culture is ordinary' (Williams 2001/1985) and that artefacts are intentional, cultural releasers 'animated by their passage through the lives of people' (Graves-Brown 2000). A material culture approach to doing history translates into an attempt at answering the question of how people have been or are by looking at what they have made. While we can debate on its scholarly value, the sword fight enacted as part of Ben Brumfield's compelling talk at Stereo Wonderland exemplified this framework in non-digital terms. My own example drew from the work of a somehow also atypical researcher, Jacqui Carey (2015) who, as part of her in depth study of the embroidered cover of a 15th century folded almanac (Wellcome Library, MS. 8932), analysed the types of silk threads being used and reproduced the spinning process. Her research focuses on the understanding of the what, why and how of past practices by re-enacting some of the making processes revealed by the artefact to her expert eyes of textile craftsperson. But what would this have to do with our engagement with digital technologies?

Digital materiality

The keynote by Bruno Latour on 'Rematerializing Humanities Thanks to Digital Traces' at the DH 2014 conference was a series of rich glosses to the statement that the digital is material. There are at least two ways in which we can grasp this. Recalling the contribution to this convention by Till Grallert around the notion of the digital being a commodity, I referred to the work of the artist Timo Arnall who recently presented his *Internet Machine* at the Big Bang Data exhibition (Somerset House, London, 2015-16; see Arnall 2014) and showcased the noisy and hulking physical materiality of digital connectivity in a film of the super secure data centre run by Telefonica in Alcalá (Spain). The digital is material also in a more subtle sense as outlined in the conversation between Laura Brake and Jim Mussell (2015). It deals with a physical, cumbersome and expensive but also *truculent*, resistant materiality with its own constraints.

Societal resonance

The theme of this convention encouraged speakers to make an explicit connection across 'Academia, Cultural Heritage, and Society.' Models used to extract patterns of significance from complex systems are ubiquitous. An attention to the materiality

of our digital world and our engagement with it seems to me of self-evident societal resonance. Only seventy years ago mainframe computers had names like *Colossus* (actually an electronic valves computer) and required expert operators to function; while nowadays it is not uncommon to see a toddler playing with a computer called *Blackberry* in her coat. If we agree with Ludmilla Jordanova (2015) on what public history is about, Digital Humanities has an important contribution to make. I presented briefly some of the works exhibited at the Big Bang data exhibition³ mentioned above which I believe resonate with our concerns as Digital Humanists, when we emphasise the importance of presentation of and interface to data as much as data collection and sampling. The importance of context and interpretation we usually make emerge via the analysis of our digital projects was revealed here through art. The contribution I see Digital Humanities making to the complex world we live in has to do with our engagement – both in teaching and research – in creating and hence also unpacking digital and data models.⁴

Modelling textuality

If the scope of material culture and the breadth of societal engagement widen our horizons with respect to the meaning of cultural artefacts and of modelling, the scope of my talk was nevertheless restricted to modelling textuality. I use ‘textuality’ on the one hand to imply a specific social theory of texts which recognises texts as open objects to be understood within the dynamic condition of their production and use (a framework articulated extensively by Jerome McGann over the years; e.g. McGann 2014); on the other hand, I use it to appeal to the readability of cultural phenomena at large (texts beyond linguistic texts). To account for this dynamic and heterogeneous view of textuality we need a model of modelling able to grasp the relational aspects of what is fundamentally a meaning-making activity (Ciula and Marras 2016). Sahle (2006; 2012) drew a very insightful model of ‘what text as reproduction is’ by plotting on a wheel diverse perspectives on textual objects useful to inform modelling efforts and in particular to develop digital editions. While there is no order in this pluralistic model of text, I would argue that an approach informed by a material culture framework would move from the upper left hand side of the wheel anticlockwise, shifting from an in depth analysis of the visual object to its semantics. In addition, such an approach would also have to consider elements outside the wheel of text, connected to the production, transmission and use of both the reproduced textual objects and the new texts being produced.

3 The curators chose the Plan of Brookes Slave Ship by William Elford (1788) to represent what they judged to be one of the first powerful data visualizations. This diagram illustrates the capacity of an 18th century ship to stack slaves; at the time it raised awareness of their inhumane treatment and had a crucial role in the abolitionist movement. Other examples taken from the exhibition I emphasised in my talk were the works ‘World Process’ by Ingo Günther (1988-) and ‘Pixellating the War Casualties in Iraq’ by Kamel Makhouloufi (2010).

4 A new project ‘Modelling between digital and humanities: thinking in practice’ funded by the Volkswagen Stiftung (2016-2018) and led by Øyvind Eide, Cristina Marras, Patrick Sahle, and myself, aims to reflect further on the kind of cultural literacy we can and want to enable via modelling.

Material primary sources (level 1) – The first level of a material culture approach to the modelling of textuality in a digital environment encompasses the materiality of primary sources, whatever our type of interest. The example I provided relates to my own PhD research (e.g. Ciula 2005), where I attempted to date and localise a certain corpus of medieval manuscripts in Caroline minuscule (X-XII centuries) drawing from the computational generation of image-like models of handwritten letters. This modelling process focused on specific features of the manuscript sources and, in particular, the morphology of the letters.⁵ However, there could be many other aspects a material culture approach to a textual source could focus on. To a certain extent this is the most obvious level of material engagement which I did not think deserved further explanation.

Materiality publications (level 2) – My second analytical level concerning the modelling of the materiality of research publications and collections we produce is possibly less obvious. The DiXiT community is certainly familiar with the notion of interface of a digital edition.⁶ Andreas Speer's talk provided rich examples of the sophisticated interfaces developed within the print tradition. The example I reflected upon draws on past co-authored research (Ciula and Lopez 2007) and focuses on the materiality of a hybrid publication resulted from the Henry III Fine Rolls project – in print, a set of volumes and on the web, a thematic research collection (Palmer 2005). Here the focus is not so much on the textual object the project departed from (13th century royal chancery documents recording fines made to the English king Henry III in exchange for favours and privileges) but on the two new publications produced by the collaborative team involved in the project.

Socio-cultural agencies (level 3) – Besides the sources we are interested in and the publications we produce, what we model is more than often a whole world around those sources and our own understanding of them. Often we call this 'data'. In the case of the project mentioned above, examples of data the historians were interested to analyse were the intersection between individuals (men and women), locations and themes or subjects. Typically, to support such analysis, it is not only the occurrence of a name in a document that needs to be recorded or encoded, but also a whole set of other information and inferences (e.g. the fact that the name refers to a person of a certain gender, engaged in a relationship of marriage with another person at a certain period as witnessed by another document etc.). The way we decided to represent such a world around the historical sources was via the development of a conceptual model or ontology (Ciula et al. 2008). At a quick glance the graph of this ontology is evidently a model which goes beyond the boundaries of the historical documents themselves to account for, for instance, a geopolitical division of England in the 13th century, the definition of what a person is, how time can be expressed etc. While one of the main purposes of creating such models might be the quantitative analysis or the compelling visualisation of certain historical information, the modelling behind it is what interests me here.

5 In Ciula and Eide (2017), we discuss more specifically how this type of image-like models come to be and what generalisations can be made with respect to inferential power and creation of new meaning

6 Pierazzo (2015) dedicates a full chapter of her book on related issues.

Whose knowledge is embedded in this model and how? What assumptions are made about the material conditions of production and use of the sources as well as of the historians' and modellers' interpretative frameworks of the same sources? These are the past and present socio-cultural agencies, some of which we model explicitly, some not.

Questionnaire to DiXiT Fellows

The rest of my talk presented the results and analysis of an online questionnaire circulated to all DiXiT fellows (13 replied i.e. ca. 86.7%). The aim of the questionnaire was to find out whether the topic I wanted to present was relevant for their research but also to challenge my own understanding and gain insights for my own ongoing research. I summarise below some of the main findings:

- There are cases where 'digital things' are by default not perceived as material in themselves.
- The majority of the fellows recognize they are engaged in heterogeneous modelling processes of some kind – mainly associated to the level 1 outlined above. Heterogeneous are also the objects being modelled (and not limited to linguistic texts).
- With respect to the work with texts, the materiality connected to the document level as well as production, transmission and use are prominent foci; however the distribution across various perspectives on text is spread out more or less equally across other levels too (including semantics). When asked to exemplify these levels, some uncertainty emerged, but again, a very rich variety of levels of attention towards the source texts was revealed.
- A surprising 38.5% of respondents do not seem to follow any specific theory of texts in their work. Yet the theories being adopted resonate with the focus of my talk.
- While the concept 'modelling textuality' might be of uncertain meaning to some (30.8%), the articulation of definitions in the responses is rich, encompassing formal and conceptual modelling as well as modelling for production (see Eide 2015).
- The research plans of the DiXiT fellowships include alternative 'products' (e.g. models and digital objects) as well as more predictable outputs (e.g. articles and monographs); the expectations of use seem to lean more towards the latter (especially articles); however the expected wide use of models is worth noticing. Whatever the expectations about use, a good 54.5% is not sure about having engaged with any foreseen uses of these products (including traditional publications).
- The examples of models being produced is very diverse ranging from informal to formal models, from prototypes to data models; same is valid for digital objects ranging from blogs to tools and digital editions (interestingly, only two of the latter have been produced as results of the fellowships though).

Some of the more extensive comments in the questionnaire were insightful in themselves; we need to think more about modelling and how we do and teach it. The analysis certainly will help me sharpen further my own research focus. I concluded summarising a research agenda for a material culture approach to modelling textuality engaged in parallel on all the three levels of modelling mentioned above. Those are the bridges we can build on to mutually relate materiality and semantics.

References

- Arnall, Timo. 2014. *Internet machine*. <http://www.elasticspace.com/2014/05/internet-machine>.
- Brake, Laurel, and James Mussell. 2015. *Digital Nineteenth-Century Serials for the Twenty-First Century: A Conversation*. 19: Interdisciplinary Studies in the Long Nineteenth Century (21). <http://www.19.bbk.ac.uk/articles/10.16995/ntn.761/DOI:10.16995/ntn.761>.
- Carey, Jacqui. 2017. 'Another enigma: reading the embroidered binding of MS. 8932.' 11 November, Wellcome library blog post. <http://blog.wellcomelibrary.org/2015/11/another-enigma-reading-the-embroidered-binding-of-ms-8932/>.
- Ciula, Arianna, and Øyvind Eide. 2017. 'Modelling in Digital Humanities: Signs in Context.' *Digital Scholarship in the Humanities* 32. suppl_1: i33–i46.
- . 2014. 'Reflections on Cultural Heritage and Digital Humanities: Modelling in Practice and Theory.' In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* 35–41. DATeCH '14. New York, NY, USA: ACM.
- Ciula, Arianna and Tamara Lopez. 2009. 'Reflecting on a dual publication: Henry III Fine Rolls print and web.' *Literary and Linguistic Computing* 24. 2: 129–141.
- Ciula, Arianna, Paul Spence, and José Miguel Vieira. 2008. 'Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project.' *Literary and Linguistic Computing* 23. 3: 311–25.
- Ciula, Arianna. 2005. 'Digital palaeography: using the digital representation of medieval script to support palaeographic analysis'. *Digital Medievalist* 1. 1. <http://www.digitalmedievalist.org/journal/1.1/ciula/>.
- Kralemann, Björn, and Claas Lattmann. 2013. 'Models as Icons: Modeling Models in the Semiotic Framework of Peirce's Theory of Signs.' *Synthese* 190. 16: 3397–3420.
- Eide, Øyvind. 2015. 'Ontologies, Data Modeling, and TEI'. *Journal of the Text Encoding Initiative* 8. <https://jtei.revues.org/1191> DOI: 10.4000/jtei.1191.
- Graves-Brown, Paul M. 2000. 'Introduction'. In *Matter, Materiality and Modern Culture*. London: Routledge.
- Günther, Ingo. 1988-. *Worldprocessor*. <http://world-processor.com/>.
- Jordanova, Ludmilla. 2015. 'A Provocation.' Public History Workshop, 29 October. <http://royalhistsoc.org/ludmilla-jordanova-public-history-workshop-a-provocation/>.
- Latour, Bruno. 2014. 'Rematerializing Humanities Thanks to Digital Traces.' Opening Plenary Lecture. Digital Humanities, July 8–12, Lausanne, Switzerland. Video available at <https://dh2014.org/videos/opening-night-bruno-latour/>.

- Makhloufi, Kamel. 2010. *Pixellating the War Casualties in Iraq*. <https://www.flickr.com/photos/melkaone/5121285002/>.
- Ciula, Arianna, and Cristina Marras. 2016. Circling around texts and language: towards “pragmatic modelling” in Digital Humanities. *Digital Humanities Quarterly*. 10, 3. <http://www.digitalhumanities.org/dhq/vol/10/3/000258/000258.html>
- McCarty, Willard. 2005. *Humanities computing*. Basingstoke: Palgrave Macmillan.
- McGann, Jerome. 2014. *New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, MA, USA: Harvard University Press.
- Palmer, Carole L. 2005. ‘Scholarly work and the shaping of digital access.’ *Journal of the American Society for Information Science and Technology* 56. 11: 1140-53.
- Pierazzo, Elena. 2015. *Digital scholarly editing: Theories, models and methods*. Aldershot: Ashgate, 2015.
- Sahle, Patrick. 2012. ‘Modeling Transcription.’ *Paper presented at the workshop Knowledge Organization and Data Modeling in the Humanities: An ongoing conversation*, Brown University, RI, 14-16 March.
- . 2006. ‘What Is Text? A Pluralistic Approach.’ In *Digital Humanities 2006 Conference Abstracts*, 188-90. Université Paris-Sorbonne: CATI, 5-9 July.
- Williams, Raymond. 2001/1958. ‘Culture is Ordinary.’ In *The Raymond Williams Reader*, edited by John Higgins. Oxford: Blackwell. 10-24.

Multimodal literacies and continuous data publishing

Une question de rythme

Claire Clivaz¹

Opening keynote given at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

The issues at stake

Fundamental questions repeatedly are raised about the evolution of the scholarly edition in a digital culture, such as the questions mentioned in the abstract of a panel organized at the first DiXiT convention in 2015 by Gioele Barabucci, Elena Spadini and Magdalena Turska:²

What constitutes the core of the edition? Its data or its presentation? In the world of printed critical editions the two things were forcefully tangled; in the electronic world they can be (and often are) separated. Is it possible to think of a digital critical edition as a collection of 'pure data'? Or is a presentation layer fundamental to the concept of 'edition'?

Such fundamental – and also disturbing – questions for the classical conception of edition matter for all the DiXiT researchers. This opening keynote lecture has presented two different issues related to them: the multimodal literacies and the continuous data publishing. A common point is used to compare them: the rhythm, according to the French expression in the title, *une question de rythme*. I wrote it voluntarily in French, first, to make visible the importance to develop multilingual perspectives in Digital Humanities; secondly, because this lecture is based on French thinkers, notably Roland Barthes and Henri Meschonnic.

1 claire.clivaz@sib.swiss.

2 See also *Digital scholarly editions. Data vs. Presentation?*, this volume.

The question of the rhythm of data publication recently has been raised strongly by a new journal in Life Sciences, *Sciencematters*,³ produced by Lawrence Rajendra, neurobiologist at the University of Zürich, and his team. He tries to advocate for publishing small data sets rather than full articles telling a ‘whole story’, but often forcing data to enter in the story, even if they are not all fitting in the narrative.⁴ The question of publication rhythm also is addressed in an SNF project developed by Sara Schulthess under my direction.⁵

Our rhythm of knowledge production is also at stake in the notion of multimodal literacies: while we are going everyday more out of the culture of the printed book, a new multimodal culture is emerging, based on the interaction between text, images, sound. And we can wonder, with Roland Barthes, if at the end the image has not always ‘the last word’ (Barthes 1995, 171). Barthes died in 1980 in a car accident, and he was preparing a seminar at the Collège de France about images, based on diapositives taken by Paul Nadar about the characters of Proust’s novel. In a 2015 monograph, Guillaume Cassegrain reconsiders all the work of Roland Barthes through the topic of the image. So if the image ‘has the last word’, are we leaving a textual civilization for a vulgar, images based culture scoop? Are we losing the fine nuances of the written text, slowly prepared in the calm office of humanist scholars, to go in the perverse ways of a simplifying oral and image based culture, that even scholars would produce as they are talking? We have surely such concerns today, the notion of rhythm is also at stake here: the rhythm with which we are thinking, talking, producing knowledge.

Multimodal literacies

My entree to this issue is one of scholarship in antiquity, a historical period where cultural transmission was primarily oral, supplemented by images and texts (Clivaz 2014). Texts were read by, at most, 10% of the population. Among these 10%, a part was able to read but not to write – a situation we will find again in the 17th century, notably, before the mass scholarization of the 19th century. Since then we have got accustomed to the equation ‘reading-writing’ that it is an effort for us today to figure what it means to be able to read without being able to write. But digital cultures and innovation soon could produce again literacies divided between reading and reading/writing. It is really impressive to see how a Google tool is now able to write what we speak.⁶ In this example, we see how ‘to talk’ is again a place of power, comparing to a time where written single literacy was dominant. In Ancient Studies, the plural has now come on the front of the scene, as shows Parker and Johnson’s collection of essays on *Ancient Literacies* (2009).

It is surely not too strong to speak today about a revenge of orality. The most advanced tool for speech recognition today seems to be the Chinese Baidu,

3 <https://www.sciencematters.io>.

4 <https://www.sciencematters.io/why-matters>; <https://humarec.org>.

5 See www.unil.ch/nt-arabe; <https://tarsian.vital-it.ch>.

6 See «Type, edit and format with your voice in Docs—no keyboard needed!», https://youtu.be/v0rPu_pl0D8; «Now You Can Edit Google Docs by Speaking», <http://www.wired.com/2016/02/now-can-type-google-docs-speaking/>

according to an announcement in the MIT Technology review: ‘China’s leading internet-search company, Baidu, has developed a voice system that can recognize English and Mandarin speech better than people, in some cases’ (Knight 2015). In a comment Andrew Ng, a former Stanford professor and Google researcher, now scientist-in-chief for the Chinese company, states that ‘historically, people viewed Chinese and English as two vastly different languages, and so there was a need to design very different features. The learning algorithms are now so general that you can just learn’ (*ibid.*).

Such an example leads us to focus on the question ‘what do we really want’, ‘what are we afraid to lose’, ‘what are we really to keep or to develop’, in our multivalent relationship with culture. I am not sure if our contemporaries are always worried about useful points. Strong reactions have been provoked in France some weeks ago by the announcement that the circumflex accent was not mandatory at school in a couple of French words (*cf.* Laurent 2016). The decision was taken 26 years ago, but has been applied only recently at school; it is still optional, homophonic words will keep it, and of course, we can still use it anyways. Nevertheless, people were so mad at this announcement by the media, that a Facebook page was opened with the sentence: ‘Je suis un accent circonflexe’, like ‘I am Charlie’, as if one was killing writing with the possible omission of the circumflex accent.

This strong protest nevertheless is mixing up written rules and living memory of orality in written rules. If we follow the work of Henri Meschonnic, there is no writing without orality embedded in it, by one way or another. Henri Meschonnic is a French thinker and writer, died 2009 in Paris, who enlightened in all his work the presence of the orality in the writing. He was a linguist and a poet and considered writing not to be opposed to orality, and sense not to be opposed to sound. Influenced notably by the reading and translation of Hebrew biblical texts, he was struck by the numerous marks of orality inscribed into the Hebrew writing. For Meschonnic, orality remains inscribed in the writing itself, and this relationship can be expressed by the word ‘rhythm’, a notion absolutely central for him: the subject who is speaking remains always related to a performance, to a social act. The speaking subject is a ‘body-social-language’. As a consequence, rhythm constitutes the principal operator of the sense in the discourse; rhythm produces the signification: ‘Le sens étant l’activité du sujet de l’énonciation, le rythme est l’organisation du sujet comme discours dans et par son discours’ (Meschonnic 1982, 217).⁷

Meschonnic’s emphasis on rhythm as organizing the discourse was anticipating the present multimodal cultural production we begin to see. Artists have been the first to explore the interactions between text, image and sound, as notably the work of the digital writer/artist François Bon shows it. His production ‘Où le monde double s’effondre’ is a particularly complicated work (Bon 2016). You have to look at it several times to grasp all the information: paintings of Pierre Ardouvin, music by Chopin, a read text by Marcel Proust, and finally, a 4th level of information built by sentences written on the screen by François Bon himself. The notion of ‘auctorial voice’, so important in all the literary studies, here is reshaped completely by the digital possibilities, offering multimodal literacies.

7 Thanks to Martine Hennard Dutheil for the reference.

Similar effects are visible also in diverse literary genres such as a photo reportage on Eretria in form of a multimedia ‘Lab’ by the Swiss newspaper *Le Temps* (Buret 2016). The rhythm of the reportage is particularly interesting: the user decides on the rhythm of images and writing, but not on the rhythm of sound. The speaking voice, most subjectively and poetically, accompanies certain images, but one cannot provoke or stop it. If one wants ‘factual’ information, as usually expected from a reportage, it is provided in written format. Communication here is deeply influenced by rhetoric.

Nowadays, the production of academic multimodal literacies is encouraged by tools like Scalar⁸ or eTalks such as those produced at the Swiss Institute of Bioinformatics.⁹ We are just beginning to face all the editorial questions raised by a multimodal knowledge: eTalks try to offer a solution to quote in detail the elements of such productions (Clivaz *et al.* 2015).

At the same time people are exploring multimodal productions and other cultural/artistic/academic productions are focusing on one sense of perception. For example, on listening and sound: The *Maison de la Poésie* in Paris proposes ‘sound snacks’,¹⁰ or artists propose to listen to a Requiem with closed eyes.¹¹ One can now consult an *Oxford Handbook of Sound Studies* (Pinch and Bijsterveld 2011). Regarding the expansion of all these elements, a module about multimodal literacies will be developed within the #dariahTeach project.¹²

Last but not least, Neal Stephenson underlined in his provocative essay ‘In the Beginning... Was the Command Line’ that orality can be found even in the code:

Windows 95 and MacOS are products, contrived by engineers in the service of specific companies. Unix, by contrast, is not so much a ‘product’ as it is a painstakingly compiled oral history of the hacker subculture. It is our Gilgamesh epic. (...) Unix is known, loved, and understood by so many hackers that it can be re-created from scratch whenever someone needs it. (...) Unix has slowly accreted around a simple kernel and acquired a kind of complexity and asymmetry that is organic, like the roots of a tree, or the branchings of a coronary artery. (Stephenson 1999)¹³

The digital culture is definitively a multimodal one – and its publication and edition remains a challenge to take up.¹⁴

Continuous data publishing

Under the label ‘stories can wait, sciences cannot’, Sciencematters defends the possibility for Life Sciences researchers to publish small datasets with comments, before having the full picture of a ‘story’ to be sold to a journal. As the website claims,

8 <http://scalar.usc.edu/about/>.

9 <https://etalk.vital-it.ch/>.

10 <http://www.maisondelapoesieparis.com/events/gouter-decoute-arte-radio-21/>.

11 <http://www.pressreader.com/switzerland/le-matin/20160515/282510067788567>.

12 In collaboration with Marianne Huang and Stef Scagliola: dariah.eu/teach.

13 Thanks to Ivan Topolsky for this reference (Kindle, l. 937-947).

14 I decided to not mention Walter Ong here, since he is quoted too often on orality.

*observations, not stories, are the pillars of good science. Today's journals however, favour story-telling over observations, and congruency over complexity. As a consequence, there is a pressure to tell only good stories. (...) The resulting non-communication of data and irreproducibility not only delays scientific progress, but also negatively affects society as a whole.*¹⁵

Changes in the rhythm of publication in Life Sciences seem mandatory today: as a matter of fact, penicillin could not get published today, as Julia Belluz (2015) claims.

If we transfer this example to Humanities, new perspectives can be perceived for digital editions: A model of continuously published small datasets need to be tested also in Humanities in close collaboration of a scientific editorial board, peer reviewers and the editors.

References

- Barthes, Roland. 1995. *Fragments d'un discours amoureux*. Paris: Seuil.
- Bon, François. 2016. *Où le monde double s'effondre*. <https://www.youtube.com/watch?v=7-IydGzDhbw>.
- Buret, Stéphanie. 2016. 'Voyage en Érythrée'. *Le Temps*. <https://labs.letemps.ch/interactive/2016/erythree/>.
- Belluz, Julia. 2015. 'Fleming's discovery of penicillin couldn't get published today. That's a huge problem.' *Vox*, 14 December. <http://www.vox.com/2015/12/14/10048422/matters-journal-small-science-big-data>.
- Clivaz, Claire. 2014. 'New Testament in a Digital Culture: A Biblaridion (Little Book) Lost in the Web?' *Journal of Religion, Media and Digital Culture* 3. 3: 20-38. Online: <https://www.jrmdc.com/journal/article/view/28>.
- Clivaz, Claire, Marion Rivoal and Martial Sankar. 2015. 'A New Platform for Editing Digital Multimedia: The eTalks'. In *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science*, edited by B. Schmidt and M. Dobrev. The authors and IOS Press. DOI: 10.3233/978-1-61499-562-3-156.
- Guillaume Cassegrain. 2015. *Roland Barthes ou l'image advenue*. Paris: Hazan.
- Johnson, William A., Holt N. Parker (eds). 2009. *Ancient Literacies. The Culture of Reading in Greece and Rome*. Oxford/New York: Oxford University Press.
- Karsky, Marie-Nadia (ed.) 2014. *Traduire le rythme. Palimpsestes* 27. Paris: Presses Nouvelle Sorbonne.
- Knight, Will. 2015. *Intelligent Machines. Baidu's Deep-Learning System Rivals People at Speech Recognition*. *MIT Technology Review*. <https://www.technologyreview.com/s/544651/baidus-deep-learning-system-rivals-people-at-speech-recognition/>.
- Laurent, Samuel. 2016. 'Non, l'accent circonflexe ne va pas disparaître.' *Le Monde*, 2 February. http://www.lemonde.fr/les-decodeurs/article/2016/02/04/non-l-accent-circonflexe-ne-va-pas-disparaitre_4859439_4355770.html.

15 <https://www.sciencematters.io/why-matters>.

- Meschonnic, Henri. 1982. *Critique du rythme. Anthropologie historique du langage*. Lagrasse: Verdier.
- . 1995. *Politique du rythme, politique du sujet*. Paris: Verdier.
- Pascal Michon. 2010. 'Rythme, langage et subjectivation selon Henri Meschonnic'. *Rhuthmos*, 15 juillet. <http://rhuthmos.eu/spip.php?article32>.
- Pinch, Trevor and Karin Bijsterveld (eds). 2011. *The Oxford Handbook of Sound Studies*. Oxford: Oxford University Press.
- Stephenson, Neal. 1999. *In the Beginning... Was the Command Line*. New York: Avon Books.

Theorizing a digital scholarly edition of *Paradise Lost*

*Richard Cunningham*¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Hans Walter Gabler defines 'a scholarly edition (a)s the presentation of a text (...) or (...) work (...) through the agency of an editor in lieu of the author of the text, or work,' but, more interestingly he notes: '(T)he scholarly edition assembles several auxiliary sections of material around the text it presents' (Gabler 2010, 44). A couple of pages deeper into that essay he clarifies that 'we *read* texts in their native *print* medium (...) in books; but we *study* texts and works in editions – in editions that live in the digital medium' (Gabler 2010, 46; emphasis added). Gabler identifies the 'auxiliary sections of material' assembled around the text as "Apparatus', 'Annotations' and 'Commentary'" (Gabler 2010, 44). The apparatus concerns itself with the material text, and it establishes the link between editor and author.

By now, 350 years after its first printing, I suggest that the editor of the next scholarly edition of John Milton's more than 10,000 line poem, *Paradise Lost*, can afford to pay comparatively little attention to the apparatus, to potential links between the editor her- or himself, and the author of the poem. As Gordon Moyles pointed out, *Paradise Lost* 'has been printed, in complete and distinct editions, at a rate equivalent to once every year' since it first appeared in 1667: probably, as Moyles suggests, 'more often (...) than any other work of literature in the English language' (Moyles 1985, ix).

Milton himself was blind by the time he composed *Paradise Lost*, and therefore no autograph manuscript has ever been, or ever could be, found. The poem's composition was enabled by the help of several amanuenses, and the only surviving hand-written portion of the poem is a 1665 manuscript in an unknown hand of Book One of the first edition. The first two printed editions of the poem,

1 richard.cunningham@acadiau.ca.

therefore, have always been regarded as the authoritative editions. The first edition was printed as a ten-book poem in 1667; the second edition was published in 1674 as a twelve-book poem. And it has been the second, twelve-book edition that, to recall Gabler's thinking, most people have *read*; but it also has been and continues to be the edition most commonly *studied*: this, despite the publication in 2007 of a new version of the 10-book first edition along with a companion volume of scholarly essays. This historical anomaly notwithstanding, the second edition remains the more often read and the more extensively studied version of *Paradise Lost*. Thus, if it makes sense to produce a new scholarly edition of the poem, it makes sense to publish a digital scholarly edition of the 1674, twelve-book, second edition. And given the impossibility of an autograph text, it obviously makes a great deal more sense to attend not to the apparatus that might link the editor to the author, but to the annotations and commentary that can be used to 'support a second 'agency' function falling to editors (...): namely, one of mediating the text, or work, and of a text's or work's meaning, to readers' (Gabler 2010, 44).

The editor of a digital scholarly edition of *Paradise Lost* will better serve the poem and the literary world by attending to the annotations and commentary that, according to Gabler, mediate the text's meaning to readers. Gabler also holds that 'annotation and commentary, are in our day widely neglected' and 'need to be brought back into' our practice as scholarly editors. But '(f)or this to come about,' Gabler continues, 'these discourses should be brought back no longer as add-ons to the edition but instead as essential strands in an edition's set of interrelated discourses, interlinked with (the text and the apparatus) not in serially additive arrangements, but in functional interdependence' (Gabler 2010, 46). That way we can realize the dynamic possibilities inherent in the digital realm, thereby better enabling *contextual* alongside *textual* study.

So that is the goal: to produce and provide Annotations and Commentaries that sit comfortably alongside, within, or in front of the text. In a paper published in volume 7 of *Digital Medievalist*, Roberto Rosselli Del Turco offers an argument that runs very nearly parallel to Gabler's. Rosselli Del Turco reminds us that '(c) omputers have become widely accepted as a research tool by Humanities Scholars in the course of the last decade. (But) academic applications like Digital Library or Digital Edition software are both limited in their diffusion and often quite complex to create.

This is especially true' wrote Rosselli Del Turco, 'in the case of Digital Edition software.' The goal of such software, he believes, 'is to integrate and interconnect several layers of information through hypertextuality and hypermediality and it faces a number of task-specific presentation problems (e.g. image-text linking, visualization of variant readings, etc.) that pose unique (User Interface) problems' (Rosselli Del Turco 2011, 3). Very soon after this point in his argument he and I part philosophical ways. Rosselli Del Turco carries on to cite Peter Robinson's 2005 'list (of) several reasons why electronic editions aren't more widespread and widely used in the scholarly community: they can be quite expensive, big publishers are not interested and, most significantly from Robinson's point of view, they're very hard to produce because we lack appropriate tools' (Rosselli Turco 2011, 3). 'To this list,' Rosselli Del Turco 'would add the difficulty from the end user perspective of

having to learn how to use a new GUI (Graphical User Interface) for almost every new electronic edition.’ It is at that point that Rosselli Del Turco and I part ways. I certainly understand his argument and I am not wholly unsympathetic toward it. It describes with a fair measure of accuracy the era of the printed scholarly edition, and in that regard arguing against it would be to argue against five centuries of successful reader-text interaction. But to wish for a one-size-fits-all software for the production of digital scholarly editions is and would be misguided, I think. I agree with Robinson’s paraphrasing of Gabler in the former’s essay ‘Editing Without Walls’ when he borrows Gabler’s phrase “dynamic contextualization” and then interprets it to signify a situation in which ‘as the reader reads through a site, all the relevant documents float to his or her attention’ (Robinson 2010, 60). But I worry that a common digital scholarly edition-producing software would deny exactly what I, Robinson, and I think also Rosselli Del Turco, want to allow: Gabler’s ‘dynamic contextualization.’

Annotation and commentary need to be brought back in to the scholarly edition. Greater attention needs to be paid to the reader through annotation and commentary than to the author through the scholarly apparatus. The result of reviving, or enlivening, annotation and commentary and re-focusing on the reader is to create, in Gabler’s words, the ‘novel opportunity of interlinked textual and contextual study’ (Gabler 2010, 46), and this is to be realized, again in Gabler’s words, through ‘the dynamics inherent in the digital medium’ (48).

So, as I theorize a digital scholarly edition of *Paradise Lost*, I suggest that a text that originates from a printed rather than an autograph source can afford to diminish the backward-looking role of the apparatus in favour of conferring greater significance on the forward- and outward-looking role of annotations and commentaries. This leads the potential editor of a digital scholarly edition of *Paradise Lost* to a position from which (s)he must consider, first, the nature of the annotations and of the commentaries, and second, the means by which those elements of the work are made available to readers. The first is a set of editorial concerns, the second poses a set of concerns to be addressed under the heading of interface. To address the editorial concerns, questions of audience must first be asked: What is meant by ‘scholarly’? Because of the nature of *Paradise Lost*, the question must be asked: how much classical knowledge do such ‘scholars’ have? how much Biblical knowledge? What, if any, specific knowledge of 17th century English history should be provided? What, if any, knowledge of the history of English Puritanism is relevant to the poem and needs to be provided? How relevant is specialized knowledge of Milton’s sui-generis species of Christianity to an understanding of the poem? To address the interface’s needs is to engage directly with the question of how to ensure the answers to the preceding editorial concerns enable readers the ‘novel opportunity of interlinked textual and contextual study’ through ‘the dynamics inherent in the digital medium’ (Gabler 2010, 46; 48).

But a diminished emphasis on the apparatus does not free one from attending to the accuracy of the digital representation of the text. And it is at that stage in which work on the digital scholarly edition of *Paradise Lost* currently is moving forward. As that work progresses, thought is being given to the preceding and other questions pertaining to the annotation, commentary, and interface design.

References

- Curran, Stuart. 2010. 'Different Demands, Different Priorities: Electronic and Print Editions.' *Literature Compass* 7. 2: 82 – 88. DOI: 10.1111/j.1741-4113.2009.00679.x.
- Dobranski, Stephen B. 2009. 'Editing Milton: The Case Against Modernization.' In *The Oxford Handbook of Milton*, edited by Nicholas McDowell and Nigel Smith, 480 – 95. Oxford: Oxford University Press.
- Gabler, Hans Walter. 2010. 'Theorizing the Digital Scholarly Edition.' *Literature Compass*. 7. 2: 43-56. DOI: 10.1111/j.1741-4113.2009.00675.x.
- Greetham, D. C., ed. 1995. *Scholarly Editing: A Guide to Research*. New York: MLA.
- Hammond, Adam. 2016. *Literature in the Digital Age*. Cambridge: Cambridge University Press.
- Hill, W. Speed. 1995. 'English Renaissance: Nondramatic Literature.' In *Scholarly Editing: A Guide to Research*, edited by D. C. Greetham. 204-30. New York: MLA.
- Milton, John. 1957. 'Paradise Lost.' In *Complete Poems and Major Prose*, edited by Merritt Y. Hughes, 173 – 469. New York: Odyssey Press.
- Milton, John. 2016a 'John Milton's Paradise Lost.' *Paradise Lost: Book I: manuscript*. The Morgan Library & Museum. Accessed November 7, 2016. <http://www.themorgan.org/collection/John-Miltons-Paradise-Lost/>
- . 2016b.'Paradise Lost.' *The John Milton Reading Room*. Accessed November 9, 2016. https://www.dartmouth.edu/~milton/reading_room/pl/book_1/text.shtml.
- Moyles, R. G. 1985. *The Text of Paradise Lost: A Study in Editorial Procedure*. Toronto: University of Toronto Press.
- Robinson, Peter. 2005. 'Current issues in making digital editions of medieval texts, or, Do electronic scholarly editions have a future?' *Digital Medievalist* 1. <http://www.digitalmedievalist.org/journal/1.1/robinson/>
- . 2010. 'Editing Without Walls.' *Literature Compass* 7. 2: 57-61. DOI:10.1111/j.1741-4113.2009.00676.x.
- Rosselli Del Turco, Roberto. 2011. 'After the editing is done: Designing a Graphic User Interface for digital editions.' *Digital Medievalist* 7. <http://www.digital-medievalist.org/journal/7/rosselliDelTurco/>
- Small, Ian, and Marcus Walsh. 1991. 'Introduction: the theory and practice of text-editing.' In *The Theory and Practice of Text-Editing: Essays in Honour of James T. Boulton*, edited by Ian Small and Marcus Walsh, 1-13. Cambridge: Cambridge University Press.
- Shillingsburg, Peter L. 1986. *Scholarly Editing in the Computer Age*. Athens, GA: University of Georgia Press.
- Tanselle, G. Thomas. *Textual Criticism and Scholarly Editing*. 1990. Reprinted 2003. Charlottesville, VA: The Bibliographical Society of the University of Virginia Press.
- Tanselle, G. Thomas. 1995. 'The Varieties of Scholarly Editing.' In *Scholarly Editing: A Guide to Research*, edited by D. C. Greetham, 9-32. New York: MLA.

The digital libraries of James Joyce and Samuel Beckett

Tom De Keyser,¹ Vincent Neyt,²

Mark Nixon³ & Dirk Van Hulle⁴

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

In 2006, Edward Beckett allowed us (Mark Nixon and Dirk Van Hulle) to consult the personal library of Samuel Beckett in his apartment in Paris and make scans of marginalia in his books. This resulted in a study called *Samuel Beckett's Library* (Cambridge UP, 2013), but we also wanted to share the digital facsimiles online. In order to do so, we needed to create a digital infrastructure. At the Centre for Manuscript Genetics in Antwerp, we already had been doing some experiments with Joyce's books and the source texts he used during the writing of his last work, *Finnegans Wake*, when it was still called 'Work in Progress'. James Joyce's writing method heavily depended on his notebooks. For his 'work in progress', he filled 50 notebooks with words and phrases, most of which were 'stolen' or 'cogged' from external source texts. In the experiment we linked the early drafts of *Finnegans Wake* to the corresponding phrases in the notebooks, and to the source texts. Whenever Joyce used a phrase in his drafts, he crossed it out so as not to use any word twice.

The notes that were *not* used are also important to genetic research. These so-called dead ends often are overlooked because no evidence of the word or phrase can be found in later stages of the writing process, but they sometimes provide useful information for the reconstruction of the writing process, especially with respect to a writer's source texts.

1 tom.dekeyser@uantwerpen.be.

2 vincent.neyt@uantwerpen.be.

3 m.nixon@reading.ac.uk.

4 dirk.vanhulle@uantwerpen.be.

Joyce's library

The first note on notebook page VI.B.6.056 presents such information. Research has shown that the word 'Kells' can be linked to *The Book of Kells*, a famous Irish medieval manuscript that contains the four gospels of the New Testament in Latin. *The Book of Kells* is an interesting source text in Joyce studies, for Joyce's paradoxical relation to his religion and nation is well known. While he rejected catholicism and Ireland, these subjects are ominously present in all of his works.

Ferrer, Deane and Lernout (2001-2004) have traced back about 35 of Joyce's notes to the introduction of a facsimile edition of *The Book of Kells*, edited by Edward Sullivan. While dead ends appear in abundance, some of the notes have been crossed out in coloured pencil and have thus been used in later stages of the writing process. We only find one such note on the first notebook page that contains references to *The Book of Kells*. This note is, along with the very first note on *The Book of Kells*, taken from the first page of the introduction by Edward Sullivan. Ferrer, Deane and Lernout (2002) have traced this note in one of Joyce's manuscripts as an addition to a fair copy of an important collection of drafts, called the 'guiltless' copybook, which comprises first drafts of several chapters of *Finnegans Wake*. The note 'spiral' became the word 'spirally' in the fair copy and that is also how it was eventually published in *Finnegans Wake* (Joyce 1939, 121.24). Interestingly, the first note – 'Kells, goldless' – is not crossed out in the notebook and does not appear in *Finnegans Wake*, but it has been transferred to another notebook, VI.C.2. At that stage, the trail stops.

To reconstruct the writing process in a genetic edition, it is important to be able to focus on the small pieces of text that Joyce was interested in. This kind of research is word or phrase based. The aim is not so much to compare different versions of a *text* (*Textfassungen* in German 'Editionswissenschaft') as it is to document the relations between the different stages of a *word* or *phrase* (or '*Wortfassungen*'; see Van Hulle 2005).

Joyce read various kinds of books, magazines or newspapers as sources of inspiration; he had some of his notebooks transcribed because of his bad eyes, and he usually worked with multiple drafts, fair copies, and typescripts before the text was sent to the printer. We can clearly distinguish different stages in the writing process. The relations between these stages define how Joyce worked, and bringing these relations together in a model reveals interesting patterns in the development of Joyce's texts.

Thanks to research projects such as the *Finnegans Wake Notebooks at Buffalo*, we now possess a large amount of data in the form of words or phrases emanating from Joyce's notebooks and manuscripts. As we are focusing on relations between parts of the data, we have brought the data together in a relational database. The database model has been designed after Joyce's own writing habits, representing 4 main stages in the writing process, namely Joyce's (1) personal library, (2) notes, (3) manuscripts, and (4) published texts.

The connections between and within the stages of the writing process are what we want to study, so we designed a system that is able to look for all links related to for instance a specific note, so as to automate what we have been doing in the example of *The Book of Kells*. The system represents the gathered bulk of

information as trees, for which each node represents a concrete stage in the writing process. A red node represents an item in the library; a yellow node refers to one of Joyce's notes, and a blue node refers to a line in a published text, in this case *Finnegans Wake*. The example of *The Book of Kells* can be generated as follows.

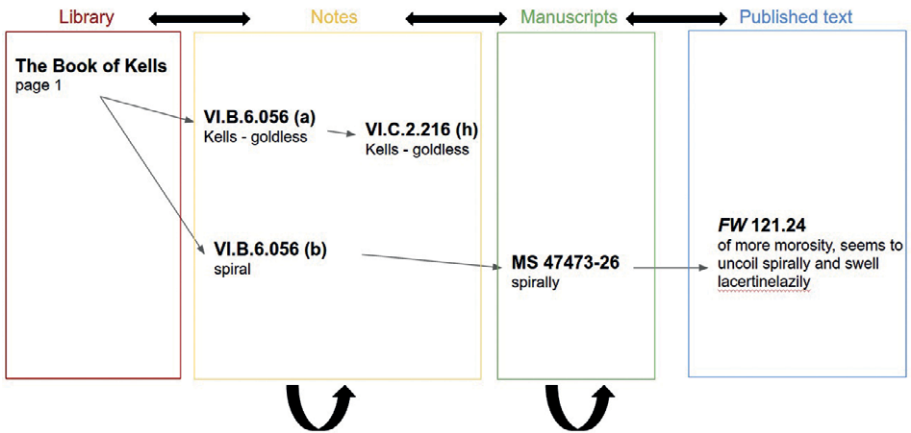


Figure 1: Visualisation of the database design based on four stages of the writing process.

Focus on links emanating from the library

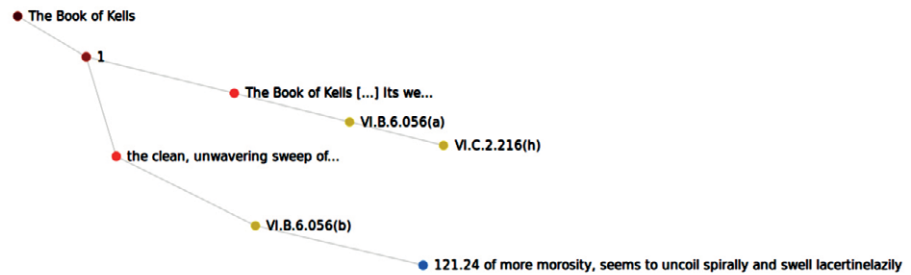


Figure 2: Writing sequence (chronological perspective) emanating from page 1 of *The Book of Kells*.

Focus on links emanating from the published text

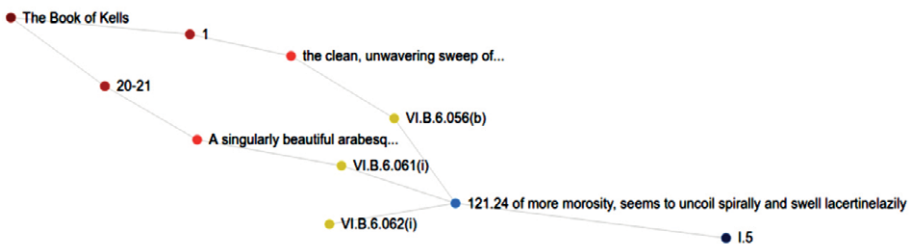


Figure 3: Writing sequence (retrograde perspective) emanating from line 121.24 of *Finnegans Wake*.

Data from research such as the *Finnegans Wake Notebooks at Buffalo* can be transferred directly into the database model. The system's design allows us to study links starting from every stage in the writing process, whether that be a source text, a note, notebook, draft, or line in a novel. Different vantage points often reveal very different information. Starting from a page from *The Book of Kells* indicates that multiple notes have been composed based on Joyce's reading of that page, and apparently one of those notes has been used for line 121.24 in *Finnegans Wake*. On the other hand, when we start searching from this line in *Finnegans Wake*, it becomes clear that two more notes have been used for its composition, which brings the total of 'used items' from notebook VI. B. 6 (for this page in *Finnegans Wake*) amount to three, two of which can be traced to *The Book of Kells*.

The digital infrastructure does not radically change the way we approach textual genetic research, but it does give shape to this research (1) in a way that enables the researcher to study the dynamics of the writing process and (2) in a way that is adapted to the nature of the data. Joyce's case suggests an approach on word or (at most) phrase level, according to a database model. Other works may require a different approach. Samuel Beckett's works, for instance, suggest an approach on sentence level, according to a text-based model.

Beckett's Library

When we started thinking about making a digital infrastructure for Samuel Beckett's library, we were more or less in a similar situation as Beckett himself. Joyce was a sort of mentor to Beckett, who helped Joyce with the correction of proofs and even the reading of books in the late stages of 'Work in Progress'. Beckett knew Joyce's system of writing from within and in his own early writing he tried to apply this system. For his first novel, *Dream of Fair to Middling Women*, he compared himself to a 'notesnatcher', referring to Shem the Penman in *Finnegans Wake*. He compiled a notebook, similar to the 50 notebooks Joyce filled for *Finnegans Wake*. But he also realized very soon that this 'notesnatching' did not work for him and that he urgently needed to find another system of writing in order to find his own voice. Initially he still had encyclopedic ambitions like Joyce and even though he later tended to present his own work as the opposite of Joyce's expansive approach, he did keep reading numerous books. Several of them are still in his apartment in Paris, even though Beckett also gave away many of his books to friends.

The library currently houses 4700 scanned pages, representing 761 *extant* volumes as well as 248 *virtual* entries for which no physical copy has been retrieved. The *virtual library* contains books we know Beckett read – because he mentions them in letters, or because they are on the TCD Calendars for the years when Beckett was a student. So far, we limited ourselves to the student library to create this virtual library. The volumes in Beckett's library show different forms of reading traces, ranging from underlined or marked passages and dog-eared pages to 'marginalia'. If the user is *only* interested in the 'marginalia', the marginalia can be arranged by the number of pages that contain annotations in the margin, which immediately shows that Proust appears twice in the Top 5, next to Maurice Scève, John Cousin's *Short Biographical Dictionary of English Literature* and Beckett's

English Bible. Especially the last volume of Proust's *Recherche* contains many marginalia, including the 'Bombardment' of involuntary memories.

Reading Proust may have helped Beckett in his attempt to take a distance from Joyce. Although he kept reading numerous books, Beckett gradually started eliminating the erudition from his works. Unlike Joyce's encyclopedism, Beckett's writing often starts from an intertextual reference, but he gradually removes it from his published texts. That means that there are sometimes intertextual references that can only be found in the manuscripts.

Suppose a user of the BDMP wonders whether Beckett used anything by her favourite poet in his works. And say that this poet is Paul Verlaine. In that case a simple search generates a survey of all the instances where Verlaine is mentioned. One item refers to a manuscript, the French draft of *Molloy*. There is no reference to Verlaine in the published work, but the manuscript refers to the last lines of one of the *Poèmes Saturniens*, a sonnet called 'Mon rêve familial' opening with the lines '**Je fais souvent ce rêve**'. It opens with a first-person 'narrator' saying that he often dreams of an unknown woman, whom he loves and who loves him ('Je fais souvent ce rêve étrange et pénétrant / D'une femme inconnue, et que j'aime, et qui m'aime'). The last stanza reads as follows (the words that are used by Beckett in the draft of *Molloy* are marked in bold typeface):

Son regard est pareil au regard des statues,

*Et, pour sa **voix**, **lointaine**, et calme, et grave, elle a*

*L'inflexion des **voix chères** qui se sont **tues**.*

The words 'voix', 'lointaine', 'chère' and 'tues' are alluded to in Beckett's manuscript, and the intertextual reference is even made explicit between brackets ('Verlaine', see Beckett 2016: BDMP4, MS-HRC-SB-4-6, 24v). But gradually Beckett hides and finally undoes the reference to Verlaine. First, in the Minuit edition, the words 'chère' and 'voix', and the name 'Verlaine' are eliminated. In the English translation, the allusion has become nothing more than a 'far whisper'. Nonetheless, the manuscript clearly indicated a reference to Verlaine. In the digital genetic edition, a note refers the user to the Pléiade edition of the Complete Works by Verlaine. Unfortunately, this is a book without marginalia, purchased long after *Molloy* was written, when Beckett finally had the money to buy it. Still, it contains an inserted card that mentions 'Je fais souvent ce rêve', indicating Beckett's continued interest in this particular poem.

This is only one particular reference to illustrate how the interaction between exogenetics and endogenetics works in Beckett's case, and how the personal library (the Beckett Digital Library) is integrated in the digital infrastructure of the genetic edition (the Beckett Digital Manuscript Project, BDMP, www.beckettarchive.org). It is only one literary allusion, found by one scholar, but a team of scholars would undoubtedly find more references, and an entire community of Beckett scholars even more. We therefore want to stress that the BDMP is a collaborative project. Wherever a reader happens to be in the Beckett Digital Library and the Beckett

Digital Manuscript Project, (s)he can make use of the button ‘Your Comments’ to send suggestions to the editorial team, for instance when (s)he has found a new intertextual reference. When the reference is added to the edition, the finder is credited automatically.

Conclusion

In non-digital scholarly editing, we usually separated the scholarly edition of an author’s texts from a catalogue of his books. Digital scholarly editing can be a great tool for genetic criticism, especially if it can include not just the endogenesis (the writing of drafts) but also the exogenesis (the links with external source texts). One way of doing this is by integrating the author’s library in the edition. Our experience is that there is not one single standard or best practice that fits all projects. The development of the digital infrastructure is determined by the particularity of each author and the extant traces of each writing process. For Joyce, we chose a relational database model, adapted to his word- or phrase-based writing. For Beckett, we adapted the digital infrastructure to a text-based approach. Consequently, there are different models of exogenetic editing (the integration of a digital library in an edition). The important thing is to embrace exogenetic editing as a new practice in digital scholarly editing; and the challenge for editors is not simply to draw attention to the links between the drafts and the source texts, but especially to develop new ways of inviting literary critics to contribute and of turning the edition into a collaborative project.

References

- Beckett, Samuel. 2016. *Molloy: a digital genetic edition* (the Beckett Digital Manuscript Project, vol. 4, BDMP4), edited by Magessa O’Reilly, Dirk Van Hulle, Pim Verhulst and Vincent Neyt. Brussels: ASP/University Press Antwerp. www.beckettarchive.org.
- Joyce, James. 1939. *Finnegans Wake*. London: Faber and Faber.
- . 2002. *The ‘Finnegans Wake’ Notebooks at Buffalo – VI.B.6*, edited by Geert Lernout, Daniel Ferrer, and Vincent Deane. Turnhout: Brepols Publishers.
- Van Hulle, Dirk. 2005. ‘The Inclusion of Paralipomena in Genetic Editions’. *Computerphilologie* 5 (23 December 2005). <http://www.computerphilologie.lmu.de/jg05/hulle.html>.
- Verlaine, Paul. ‘Mon rêve familial’ (from: *Poèmes saturniens*).

Editing the medical recipes in the Glasgow University Library Ferguson Collection

Isabel de la Cruz-Cabanillas¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Introduction

Medical recipes used to be incorporated into other codices which contained more extensive and relevant works. Therefore, many recipes have remained hitherto unknown and the only way to identify them is by consulting different catalogues and manuscripts. Nonetheless, even specialised catalogues are rarely comprehensive and do not include cross-references to other catalogues, which makes the identification of recipes an arduous task and, consequently, their edition and study.

This paper reports on a project on the John Ferguson Collection of medical recipes carried out at Glasgow University Library whose aim firstly was to identify English medical recipes in the Ferguson Collection and secondly to edit them in order to undertake the study of their linguistic and structural features and analyse their development over time.

John Ferguson was a Chemistry Professor at Glasgow University from 1874 to 1915. Ferguson's personal library was extensive, containing approximately 18,000 volumes. After his death, an important part of his collection was purchased by Glasgow University in 1921. The collection is made up of c. 500 manuscripts and c. 7,500 printed books from 1363 to 1864 and is written in different languages, chiefly in Latin, German and English, but there are also texts in Italian, French and Portuguese. The manuscripts in the collection are mainly about chemistry, alchemy and medicine. The library catalogue, as well as other specific catalogues on medical and scientific manuscripts, were searched to select the manuscripts

¹ isabel.cruz@uah.es.

that should be scrutinised (Keiser, 1998; Ker, 1977; Voigts and Kurtz, 2000). The compilation of previously unexplored English medical recipes in the Ferguson Collection is structured in the following way:

- a. Alchemy treatises (Ferguson MS 58, Ferguson MS 91, Ferguson MS 229 and Ferguson MS 309), which contain a small number of medical and culinary recipes.
- b. Medical compendia, such as Ferguson MS 147, where the recipe collection appears along with other more well-known treatises, as it is the case of the *Antidotarium Nicholai*.
- c. Recipe books, like Ferguson MS 61, Ferguson MS 15 and Ferguson MS 43, where medical, cooking and other kinds of recipes are bound together in one single volume.

In fact, Taavitsainen (2009, 194) noticed that ‘texts from one genre, such as recipes, can occur in several traditions.’ Thus, as can be deduced by the groups above, recipes appear in medical compendia, but also in collections including cooking recipes, as noted by Görlach (2004), and alchemical books.

Medical recipes in the Ferguson Collection

Alchemy Treatises

The recipes in this section are scarce, ranging from only one recipe in Ferguson MS 91 to several in Ferguson MS 58. They are all alchemy treatises from the 16th and 17th centuries dealing with works by George Ripley and other well-known alchemists like Raimundus Lullius, where medical recipes are included along with alchemy recipes, as in Ferguson MS 229 and Ferguson MS 309.

Medical Compendia

Ferguson MS 147 proved to be one of the most interesting pieces for the project, as it contains a wide collection of medieval recipes in English as well as some charms in Latin. A specific challenge was the transcription of shortenings (especially a superscript 9 for *-us* and the shortening for *-ur*). The superscript 9 seemed to work perfectly in Latin words, but produced a weird type of language in English native words, with forms like *baronus*, *onus*, *ranckelus* or *hornus*, reinforced by other expanded forms like *shepus*, *skynus*, *gostus*, *cornus*, *monthus*, *clothus*, *clessus*, *cropus*, or *bretecropus*. Likewise, the expansion of the shortening *-ur* fit well in Latin words, such as in *sanabitur*, but incorporated a lot of native forms like *watur*, *botur*, *powdur* or *togedur*. Nonetheless, the study of these items along with other dialectally significant forms, according to the grounds established by the LALME team, led to the localisation of the language of the Ferguson MS 147 in an area delimited by Shropshire to the north, Herefordshire to the south and Monmouth to the west (De la Cruz-Cabanillas, 2017b).

Regarding the structure of the recipe the usual elements are found: The title or medical purpose introduced by *Medycin for/to do*, *To make*, *Another for the same* or *Another*. This is followed by the ingredients section, where usually plants combined with the juice of fruits or other liquids, such as wine or water, are needed to prepare

the recipe. Sometimes sugar or honey may be added and some other ingredients as well. After that, the reader faces the preparation section with instruction in relation to the combination of ingredients. In the preparation phase, culinary verbs in the imperative mood, as well as technical vocabulary related to kitchen and medical utensils are found often. Then, the application section presents a less well-defined organisation of information. It describes how the remedy is to be used by indicating its dosage and duration. Finally, the efficacy phrase evaluates the recipe with an English sentence or the Latin formula *sanabitur* (see De la Cruz-Cabanillas, 2017a).

Recipe Books

The three recipe books in the Ferguson collection were compiled by women in the 17th and 18th centuries. They depict women who are aware of their role in the household and the community, which consists of caring about people's health in their environment. For this reason, women gather information relevant to this purpose and any other recipes that may be useful to them. Thus, Mary Harrison (Ferguson MS 61) includes not only medical recipes but also recipes for cattle, chicken or instructions on how to polish men's boots properly. Her few cooking recipes have a therapeutic purpose, broths are recommended for strengthening and the recipe for 'Pepper Cakes' is followed by a section which specifies the virtues or 'uses of it', where the cakes are claimed to be good for digestion, as well as for the brain and to restore your memory (See De la Cruz-Cabanillas, 2016). Similarly, Ferguson MS 43 includes mainly medical recipes with one or two cooking recipes, unlike Ferguson MS 15 whose main contents deal with cooking instructions and just the final part is devoted to a brief section on medical recipes.

English recipes across time: from Late Middle English to the 18th Century

If the focus is now on how the structure of the recipe has evolved, as noted by Jucker and Taavitsainen (2013, 147), 'genres show different realisations in different periods, but more prototypical features may remain constant in a long diachronic perspective.' This statement holds true for the Ferguson medical recipe collection, since the same structural constituents are preserved from late Middle English up to the 18th century, where slight differences are observed.

In the title section, the same linguistic procedures which were available in the Middle English period are registered in Early Modern English, as can be seen in Table 1.

Regarding the ingredients section, *take* is still the prevailing verb, but other imperative forms also appear. It is worth noting that new ingredients, such as tea, coffee, chocolate or sarsaparilla, are now available to be used in preparing recipes. This section is followed by the preparation section characterized by a gradual specialisation conveyed by verbs such as *boil*, *seethe*, *mix* or *strain*. In the application section, a major specialisation can also be observed with indication of exact measures, although vagueness is still present, as when the reader is asked to take 'as much green young parsley as you can hold betwixt yr four fingers' (Ferguson MS 15). Finally, the efficacy phrase is not so common in the Ferguson

Template	Example
<i>To + inf. + NP</i>	To make Surfeit Water (Ferguson MS 15) To make ye White Salve excellent for all Wounds & Aches (Ferguson MS 43)
<i>For to + inf.</i>	For to make Tincture of Carroways (Ferguson MS 15)
<i>For + v-ing</i>	For pising a bed (GUL Ferguson MS 61)
<i>For + NP</i>	ffor ye bloody flux (Ferguson MS 229) For ye Wormes (Ferguson MS 43)
<i>NP + for + NP</i>	An Excelent Plaster for A Cosumption (Ferguson MS 61) A Plaister for an Ache (Ferguson 43)
<i>NP + to + inf. + NP</i>	A Remedye to break the stone (Ferguson MS 58) A Water to stay ye fflux (Ferguson MS 43)
<i>NP</i>	dr Stephens water (Ferguson MS 15) A Purging and clensing Julep (Ferguson 43)
<i>Another (+NP)</i>	an other (Ferguson MS 61)
<i>Another for the same</i>	An other Couling Ojntment for ye Same Vseies (Ferguson MS 61)

Table 1: Formulaic templates in recipe titles.

collection of Early Modern medical recipes, as it used to be in medieval times. In those cases where it is present, the Latin set phrase *sanabitur* is no longer used.

Conclusions

The research presented here is part of a project investigating the extensive material in the collection of John Ferguson housed at Glasgow University Library. The interest in this specific collection lies in the fact that it contains unexplored texts, difficult to spot, since they are not acknowledged properly in catalogues and must be identified by searching through the manuscript pages.

The medical recipes in the collection have been identified in different genres which comprise alchemy treatises, to medical compendia or recipe books containing not only medical but also cooking and practical recipes to manage Early Modern England households.

Regarding the analysis of the genre and text type conventions of the recipes, it has been observed that the main constituents of the recipe structure remain from medieval times; namely, the title to show the purpose, the ingredients section, the preparation of the recipe and a possible final phrase to evaluate its efficacy. Regarding the linguistic formulation of the latter section, it is observed that the Latin phrase *sanabitur*, found in medieval recipes, has disappeared completely in the Early Modern English period.

References

- Michael Benskin, Margaret Laing, Vasilis Karaiskos, and Keith Williamson. 2013. *An Electronic Version of A Linguistic Atlas of Late Mediaeval English*. Accessed, July, 2016. www.lel.ed.ac.uk/ihd/elalme/elalme.html.
- Isabel De la Cruz-Cabanillas. 2016. 'Mary Harrison's Book of Recipes. Women and Household Medicine in Late 17th Century.' *Revista Canaria de Estudios Ingleses* 72. 1:79-95.
- Isabel De la Cruz-Cabanillas. 2017a. 'Medical Recipes in Glasgow University Library Manuscript Ferguson 147.' In *Essays and Studies in Middle English*, edited by Jacek Fisiak. 77-94. Frankfurt am Main: Peter Lang.
- Isabel De la Cruz-Cabanillas. 2017b. 'Mapping the Language of Glasgow University Library Manuscript Ferguson 147.' In *Textual Reception and Cultural Debate in Medieval English Studies*, edited by M^a José Esteve-Ramos, and José Ramón Prado Pérez. Cambridge Scholars: Newcastle upon Tyne.
- Görlach, Manfred. 2004. *Text types and the history of English*. Berlin: Walter de Gruyter.
- Glasgow University Library Catalogue: Ferguson Collection*. Accessed, July, 2016. http://special.lib.gla.ac.uk/manuscripts/search/detail_c.cfm?ID=70.
- Jucker, H. Andreas, and Irma Taavitsainen. 2013. *English Historical Pragmatics*. Edinburgh: Edinburgh University Press.
- Keiser, George R. 1998. Works of Science and Information. General editor. Albert E. Hartung. *A Manual of the Writings in Middle English 1050-1500*. Vol. 10. New Haven, Conn.: The Connecticut Academy of Arts and Sciences.
- Ker, Neil R. 1977. *Medieval Manuscripts in British Libraries*. Oxford: OUP. 5 vols.
- Taavitsainen, Irma. 2009. 'Early English Scientific Writing: New Corpora, New Approaches.' In *Textual Healing: Studies in Medieval English Medical, Scientific and Technical Texts*, edited by Javier E. Díaz-Vera, and Rosario Caballero, 177-206. Bern: Peter Lang.
- Voigts, Linda Ehram, and Patricia Deery Kurtz. 2000. *Scientific and Medical Writings in Old and Middle English: An Electronic Reference*. Ann Arbor: University of Michigan Press.

The archival impulse and the editorial impulse

*Paul Eggert*¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The distinction between archive and edition has been much discussed in recent years. Some people believe the distinction ought to be a firm one. In an article from 2013, for instance, three scholars associated with the *FaustEdition* read the archive versus edition distinction back onto Hans Zeller's famous distinction of 1971 between *Befund* and *Deutung* (record and meaning) (Brüning *et al.* 2013; Zeller 1995). Although I too have been attracted to the parallel I now feel that applying Zeller's case in this way is a mistake. Unfortunately the two categories – record and interpretation – cannot gain the firmly differentiated objective footing for which philologists traditionally have yearned. This is because humanly-agented reading is intrinsic to both of them. There is no firm, outside vantage-point from which to survey and thus to define archive and edition as securely differentiated categories. As readers we inhabit the same textual field as the approaches to documents and texts that we seek to define. To *record* is first to read and analyse sufficiently for the archival purpose; to *interpret* is first to read and to analyse sufficiently for the editorial purpose. In practice, the archival impulse anticipates the editorial, and the editorial rests on the archival. They are not separate categories and certainly not objective or transcendental ones. As co-dependents they are perhaps best understood as being in a negative dialectical relationship with one another, that is, each requiring the other to exist at all, yet each requiring the other's different identity to secure its own.

So, how better may we envisage the relationship of archive and edition? I propose that we think of a horizontal slider or scroll bar running from archive at the left to edition at the right. In this model every position on the slider involves interpretative judgement appropriate to its purpose. All stakeholders will want to

¹ pauleggert7@gmail.com.

position themselves at one place rather than another along that slider, or at one place rather than another depending on what responsibilities they are discharging at any one time. To see things like this is to want to recast the archive idea as archival impulse and its complementary opposite as the editorial impulse. Based on impulse or tendency, the distinction that I am proposing is pragmatic rather than ideal. To the extent that the archive – edition distinction is deployed it has to be understood, then, as a shorthand standing in for what is in fact a more nuanced reality. I wish now to consider where this approach leads us.

Every position along the slider involves a report on the documents, but the archival impulse is more document-facing and the editorial is, relatively speaking, more audience-facing. Yet each activity, if it be a truly scholarly activity, depends upon or anticipates the need for its complementary or co-dependent Other. The archival impulse aims to satisfy the shared need for a reliable record of the documentary evidence; the editorial impulse to further interpret it, with the aim of reorienting it towards known or envisaged audiences and by taking their anticipated needs into account. Another way of putting this is to say that every expression of the archival impulse is to some extent editorial, and that every expression of the editorial impulse is to some extent archival. Their difference lies in the fact that they situate themselves at different positions on the slider.

The slider model permits certain clarifications to emerge about archival – editorial projects in the digital domain. On the very left of the slider where the archival impulse is dominant, there is, to be strict, no robust or settled work attribution yet available, nor for that matter a version concept, since, strictly speaking, all the archivist-transcriber has are documents in need of transcription. Work-and-version attributions are a matter for editorial postulation and argument once the facts are in, once the record has been more or less settled. In practice of course, work-and-version concepts often are drawn down in advance, whether from tradition, from the fact of preceding publication or following bibliographical or codicological analysis. The terms ‘work’ and ‘version’ are useful categories by which to organize the archival effort, to keep it within achievable bounds. Their use is a silent indication that the archival impulse already bends, if ever so slightly, towards the later editorial one. Thus the two are linked, because they are in need of one another, even at this early stage. Nevertheless, the drawing-down has to be understood as provisional and always open to editorial reinterpretation or challenge.

As the project subsequently proceeds, the document-facing transcription, now tentatively completed, begins to come into a dawning editionhood of its own as the scholar-transcriber draws the documentary details into a shaping editorial argument about their history and significance. Even if originally prepared for private uses, the transcription now bends that document’s textual witness towards an envisaged readership. Any and every emendation that makes the text more legible or usable is done on behalf of a readership, and that fact shifts the project a few points along the slider to the right without having quite reached the midpoint. Nevertheless, the archival pull of the document, of fidelity *to* the document, remains strong. That is the first clarification.

Once the relevant transcriptions for a multi-witness work have been prepared and checked, once the archival impulse has been satisfied, a second clarification emerges from the slider model: automatic collation is the pivot between the archival impulse and the editorial impulse. It is the midpoint on the slider. The editor considers the meanings of the data produced by the collation and confirms, corrects or discards the provisional assignation of versionhood and workhood that had helped organize the archival effort. The editor finds it harder and harder to resist the pull of work or version concepts as containers for shaping the data into a legible form for readers. As the editorial impulse gains ascendancy – as archival data is converted into evidence in support of the editorial argument – a more fully reader-facing edition comes into focus. Documentary fidelity is by no means lost sight of – the slider model insists on it – but is now consigned to the archival expression of the project.

Although the slider links all interpretative archival – editorial decisions on the same continuous scale it is obvious that a transcriber's decision to record as unobjectionably as possible the attribute and rend tags for, say, italics involves a different level of judgement than the system-wide decisions that an editor must make. The project workflow wisely will respect that reality. It normally will make sense to do the reader-oriented editing after the archival effort is finished, even though, admittedly, the archival phase will be generating all kinds of clues that will benefit the editorial effort.

But what form should the digital edition take and how should it be stored, joined at the hip as it is to the archive from which it now seeks separation? Or, put another way, as it reaches towards the right-hand end of the slider – for a reader-facing editionhood? So far we know that such a digital edition typically will take the form of a reading text of the unit being edited (work or version) or the genetic development of the version or draft, supported by a commentary analysing the documentary evidence. The edition will be potentially only one considered application of that data, potentially only one of many, since there will usually be more than one possible argument about textual authority or authorization. Indeed, there will potentially be as many editions as there are organizing arguments. Each one, offered as an interpretation of the data in the archive, must then take its chances in the intellectual marketplace. It must persuade its readers or, in the case of a genetic edition, be useful to those who study the genesis or emergence of text on the page under the hand of the writer.

The sliding scroll-bar model dispenses with anxiety about archives replacing editions. Editions will continue to be prepared as long as there are readers whose requirements and capacities need to be served. Readers need reader-facing editions. The special-purpose collections that we call archives can be wonderful achievements in themselves, and certainly they are indispensable to digital editing. But only a tiny number of readers will want or feel that they need face-to-face engagement with the original sources or will be able to make effective use of them. Expanding the constituency of such curious readers to help them engage with the primary documents is desirable, for a range of pedagogic reasons. But we have to be realistic about our chances of success. Accordingly, I believe we will continue to prepare editions in the digital domain if only because, on the slider, archival contributions

are by definition document-facing. They have a responsibility to leave the textual difficulties encountered in situ for the next specialist user. Ordinary readers, on the other hand, will need those impediments to comprehension to have been digested editorially by commentary or emendation.

Because of the way in which it will be stored the digital edition will be better described as the editorial layer of the complete project. It will be the normal point of entry for the reader. It will be up to the editor to link the archival evidence to the editorial layer again and again, thus tempting the reader to go deeper. Provided that the project can be given a collaborative interface some readers may become enduring contributors to the project. This is the third clarification: apprehensions that single-document transcription projects are replacing work editions needlessly telescope the slider into a single disputed point. The slider model helps us to survey the full range of archival and editorial possibilities.

Patrick Sahle's recent use of the term 'critical representation' to describe the range of project outcomes between archive and edition also warrants inspection. He writes: 'A scholarly edition is the *critical representation* of historic documents' (Sahle 2016, 23; emphasis in original). The benefit of the term 'critical' lies in the fact that it allows him to envisage archival and editorial endeavours as more or less continuous without sacrificing any of the rigour associated with scholarly editing. But it does not afford a way of differentiating expressions of archival from editorial impulses such as I have been arguing is essential.

The term 'representation' seems to apply aptly to the aim of transcribers of documents, especially in projects involving the encoding of the transcribed text. So the term embraces the archival impulse nicely. But it does not satisfactorily describe the editorial impulse. The editor's aim is less to *represent* something that is pre-existing than to *present* something (the text of a version, the text of work, a process of writing) that typically has not existed in precisely this form before, together with the critically analysed materials necessary to defend the presentation.

While an archival transcription is an attempt to capture the text of a historical document (representation), an edition claims to make present the text of the thing that has been subject to the editorial analysis (presentation). This distinction seems cleaner – there is less muddying of the waters – than trying to bridge, as Sahle does, the archival and editorial activities as both being forms of representation.

References

- Brüning, Gerrit, Katrin Henzel and Dietmar Pravida. 2013. 'Multiple Encoding in Genetic Editions: The Case of Faust', *Journal of the Text Encoding Initiative*, 4 (March).
- Sahle, Patrick. 2016. 'What Is a Scholarly Digital Edition.' In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo. Cambridge, UK: Open Book Publishers.
- Zeller, Hans. 1995 (originally in German, 1971). 'Record and Interpretation: Analysis and Documentation as Goal and Method of Editing.' In *Contemporary German Editorial Theory* edited by Hans Walter Gabler, George Bornstein and Gillian Borland Pierce, 17-58. Ann Arbor: University of Michigan Press.

Pessoa's editorial projects and publications

The digital edition as a multiple form of textual criticism

Ulrike Henny-Krahmer¹ & Pedro Sepúlveda²

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

A digital edition of Fernando Pessoa's editorial projects and publications is being established through a collaboration between scholars from the Institute of Literature and Tradition (IELT) of the New University of Lisbon and the Cologne Center for eHumanities (CCeH) of the University of Cologne.³ The edition focuses on the contrast between the potential character of Pessoa's numerous lists of editorial projects and his actual publications in lifetime. The digital format allows for a combination of different editorial procedures, proposing a new form of a multiple textual criticism within the editing of Pessoa's work, where coexistent forms of transcription are applied together with genetic and semantic criticism.

Pessoa's editorial projects and publications

Fernando Pessoa conceived numerous lists of editorial projects, showing a constant obsession with the editorial planning of his work. These lists follow not only an editorial purpose, foreseeing the future publication of the work, but define also systemic grounds of this work. By establishing titles and attributing them to

1 ulrike.henny@uni-wuerzburg.de.

2 psepulveda@fcsh.unl.pt.

3 The project has been primarily funded by the Fundação para a Ciência e Tecnologia (FCT) and additionally by the CCeH to fund transcribers, encoders and programmers, mainly between September 2014 and June 2015. A special thank you goes to Ben Bigalke who presented the poster 'Digital Edition of Fernando Pessoa: projects & publications' at the DiXiT convention in Cologne.

certain author names or to collections of works, the lists of editorial projects give meaning to otherwise loose texts. The great amount of lists which remained in Pessoa's Archive, held by the Portuguese National Library (BNP), and some still in possession of the heirs, are of decisive importance for the understanding of the foundations of Pessoa's work and its development. The significance of this *corpus* questions the idea of a poet in search for anonymity and avoiding the publication of his work. Although publishing little in lifetime, at least if we compare these publications with his Archive, gathering almost 30,000 documents, Pessoa was an author for whom the edition and publication of his works was of utmost importance. Parallel to his writings, the poet constantly developed lists gathering editorial projects, providing his work with a potential dimension, exceeding what was in fact written or even published.

As the lists show the development of titles, authorial attributions, and the establishment of collections of works throughout different periods of time, they allow to chart the history of Pessoa's projects, regarding each particular work. The lists of editorial projects are to be distinguished, within Pessoa's Archive, from other types of lists, such as lists of tasks or of readings, and from editorial plans, that provide the structure of one particular work. These lists refer to editorial projects of a diverse nature, establishing relations between them. The simultaneity of the occurrences of certain projects, in relation to others, is decisive for the understanding of the history of their development.

Although the importance of the editorial lists has been pointed out by recent criticism (*cf.* Sepúlveda 2013 and Feijó 2015) and they have been included in several editorial volumes, a comprehensive edition of these lists is still missing. Following a first gathering of these documents by Jorge Nemésio in 1958, Teresa Sobral Cunha referred in 1987 to the plan of establishing an edition of these lists, among which the plans structuring the different works, without having ever concluded the task. Never having been conceived to be published, even less integrated in a book, as these lists represent plans for future publications, the dynamics of the digital provide the adequate support for their publication. The *corpus* of the edition will include, in a first launch, to be concluded in 2017, all lists of editorial projects which were located in the Archive held by the Portuguese National Library, as well as by the heirs, elaborated between 1913 and 1935. In addition to this *corpus*, Fernando Pessoa's published poetry in lifetime, in journals and literary reviews, between 1914 and 1935, also will be integrated in the digital edition, already in its first launch. 1914 is the year of Pessoa's first publications of poetry in Portuguese until his death in November 1935, a period in which Pessoa publishes his poetry written under the fictional names of Alberto Caeiro, Álvaro de Campos and Ricardo Reis, defined by the poet as heteronyms, as well as a significant part under his own name. The extension of the *corpus* to the poetry published in lifetime allows for an interaction between the potential level of the projects and what has been presented effectively in terms of publication. This interaction is facilitated through the establishment of links between the documents, of different indexes and a timeline of all lists and publications included.

Multiplicity and simultaneity of textual criticism in the digital edition

Coexistent forms of transcription

The TEI-based digital edition offers to the reader a combination of different editorial modalities for each document of the Archive, thereby proposing a new form of textual criticism within the edition of Pessoa's work. For each document, four different forms of transcription are presented:

- A diplomatic transcription, including all variants, hesitations and passages later rejected by the author.
- A first version of the text, as established by the author, including the development of abbreviations.
- A last version of the text, following the last non-rejected textual variant, also including developed abbreviations.
- A personal version of the text, allowing for the reader to establish his own version, by choosing among the elements he wishes to see presented.

Due to the existence of several textual variants within the archival writings, previous editions, in a book form, establish only one particular version of the text, by choosing among the variants. Some editions, such as the collection of works published by Assírio & Alvim, define themselves by choosing the first version of each text, based on the argument that further variants are mere hesitations of the author (*cf.* Lopes 1992). Other editions, such as the critical edition of Fernando Pessoa, published by Imprensa Nacional-Casa da Moeda, follow the last textual variant, arguing that this corresponds to the poet's last intention (*cf.* Duarte 1988 and Castro 2013). Even further editions choose a hermeneutical criterion as the basis of their editorial proposal, by choosing among the variants the editor sees as the most adequate in a given text (*cf.* Zenith 2011).

This digital edition offers the different modalities of reading each text, by not rejecting any of the textual variants left by the author in his archival writings, as it is the case with the lists of editorial projects. This editorial proposal, facilitated by the digital format, provides a privileged access to the author's writing process, offering the reader the different textual versions, corresponding to alternatives and hesitations left by the author. The textual variants, the underlined, as well as the rejected passages are presented in terms of a graphical proximity to the textual source. A facsimile of the text is visible beside each transcription.

Genetic and semantic criticism

The approach to offer coexistent textual layers in an edition, not to prefer one authorial textual variant over another, and not to establish a definite and unambiguous version of a text meant to be the final and authoritative one, has been facilitated by the digital medium. Notwithstanding the novelty of this procedure inside the tradition of editing Fernando Pessoa's works, it has been practiced elsewhere. As Elena Pierazzo points out, the multiplicity of transcriptions in a

single edition may be due to a by now established model for digital editions which she calls the ‘Source-and-the-Output-Model’ (2015, 25):

most digital scholarly editions are based on markup (...) This fact then implies that we need to distinguish the data model, where the information is added (the source), from the publication where the information is displayed (the output) (...) type-facsimile or diplomatic, or (...) so-called reading editions (...) each represents only one of the possible outputs of the source.

She then proposes that the classic labels for editions should be redefined. In fact, when thinking about how the digital edition of Pessoa’s editorial projects and publications could be classified according to the editorial procedures pursued, we found that it is best described as an edition making use of various, multiple forms of textual criticism, borrowing from and combining different editorial approaches. Following the basic typology proposed by the Institute for Documentology and Scholarly Editing (IDE) in a questionnaire which has the purpose of gathering information about a variety of digital editions, one could say that the Pessoa edition includes at least aspects from a documentary, a diplomatic, a genetic and an enriched edition:

Documentary Edition: Related to the school of ‘documentary editing’, focuses on the documents rather than on abstract texts; tries to give truthful representations of the documents with (often: diplomatic) transcription and additional information.

Diplomatic Edition: Focuses on the text (not the visual layer) of documents, tries to give a transcription as accurate as possible.

Genetic Edition: Focuses on the genesis of texts, usually on the level of microgenesis (within a document) sometimes on the level of macrogenesis (across documents). (...)

Enriched Edition: ‘Enriched Edition’ describes digital representations of texts that put a particular emphasis on extracting information from the text, e.g. by elaborate indices, extensive comments and annotations on the content, linking of related information, or formal representation of content.

(IDE 2014ff.)

Another matter in the questionnaire regards the forms of text which are presented in an edition. From the six different options offered (Facsimiles, Diplomatic transcription, Edited text, Translations, Commentaries, Semantic data), five apply to the digital edition at hand (all except translations). Luckily and probably significantly, most of the questions allow for multiple choices of answers. Two of the above mentioned facets are elaborated a bit further here to show in what way they apply to the edition of Pessoa’s projects and publications, namely semantic and genetic criticism.

Genetic editions focus on single documents, the relationships between documents, the process of a text coming into being and the evolution of a work (cf. Deppman *et al.* 2004; Gabler 1999). In this respect, the digital edition of Fernando Pessoa’s projects and publications can be seen in the context of genetic

criticism: single handwritten or typed documents are transcribed, thereby capturing additions of characters, words, lines, marginal notes; substitutions and deletions made by the author. The goal is to trace the evolution of Pessoa's work as planned and projected by himself in the editorial lists. Therefore, the documentary and genetic approach are not followed exhaustively:

- the transcription is made only for those parts of the documents that contain editorial lists and poems published by Pessoa during lifetime and are thus of interest for this edition
- additions, substitutions and deletions are only recorded if they are meaningful, in the sense that they cause changes to author names, work titles etc.

Second, the notion of genesis assumed in the edition needs a closer look. What is of interest here is not a single specific text or work whose genesis is traced from a first sketch on one or several documents to a fully elaborated version. Rather, the focus is on what could be called a meta level. The documents witness mentions of author names and titles in editorial lists, first of all, they 'speak about' and evoke a work (cf. Figure 1).

The material documents point to an abstract notion of work, whose title, (fictitious) author and structure can change, regardless of whether the work materializes at some point or does not. The meta level is supposed to mean that changes to an abstract text or work are *described* by what has been written in the documents instead of being *manifest* in the writing.

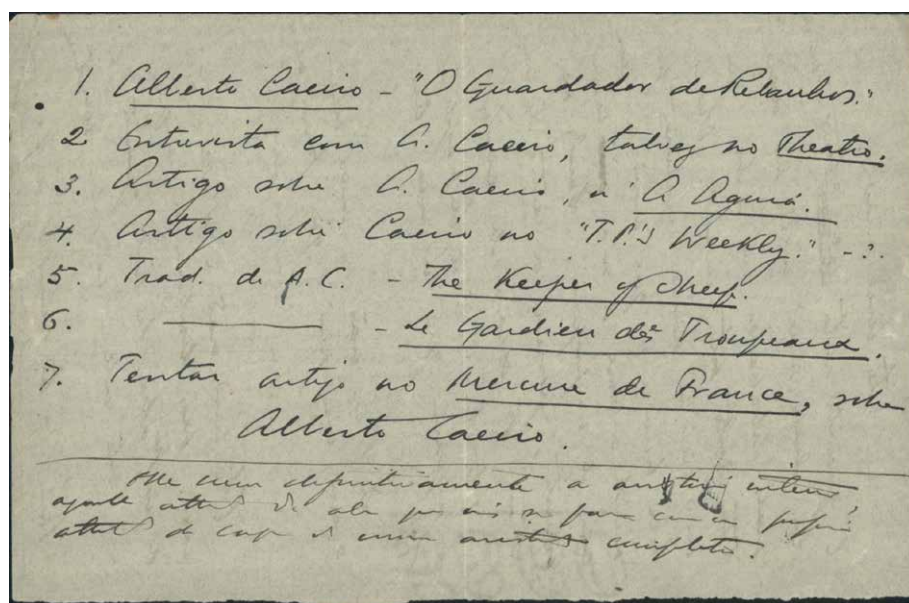


Figure 1: Example of a document (BNP/E3 8-3v) from Pessoa's archive, containing a list of planned texts and works.

A change of a work's title can for example occur when:

- a work title is mentioned in an editorial list on one document and there is an addition, substitution or deletion of a word of the title on the same document
- a work title is mentioned in an editorial list on one document and mentioned again on another document, but with a change.

That way, the partially imagined work of Fernando Pessoa emerges from the documents which contain the editorial lists, plans and notes, as sketched in Figure 2.

To be able to trace and analyze the evolution of Pessoa's work as planned by himself, it is necessary to know when references to author names refer to the same person and references to work titles refer to the same work. This is where semantic criticism comes in. The encoding of semantic information has played a role in the context of digital editions and the TEI.⁴

For the purpose of identifying persons (in the role of authors, editors, translators, topics) and works, a central list of imagined and historical persons and works has been established in the project. References in the documents are encoded and point to that central list. Obviously, it is not always clear to what work a title refers, especially when the title can undergo a change. It has been decided to start with the cases that are less doubtful, in order to be able to draw conclusions from the semantic encoding. Figure 3 shows an RDF graph as an example of a formal description of the relationships between works, work titles and documents. The work O1 ('Poems of Fernando Pessoa') can have the title 'Cancioneiro' as well as 'Itinerario'. In the document BNP/E3 63-31 the title 'Itinerario' is mentioned and thus the work O1.

An analysis of a handful of already encoded documents from 1915 to 1935 with references to the 'Poems of Fernando Pessoa' shows that Pessoa used the title 'Cancioneiro' and 'Itinerario' alternatively in his editorial lists. A slight preference for 'Itinerario' in the early documents and for 'Cancioneiro' in the later ones becomes visible and might be supported or rejected by further documents. Interestingly, the different titles are even used in the same document, *e.g.* BNP/

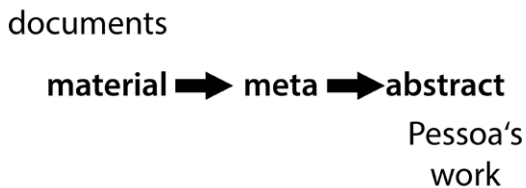


Figure 2: Pessoa's (abstract and partially imagined) work emerging from the documents with editorial projects.

4 *cf.* Eide 2015 on the relationships between the TEI and ontologies and MEDEA, a project focussing on the development of standards for the encoding of semantically enriched digital editions of accounts, *cf.* <https://medea.hypotheses.org/>.

E3 44-47r. In the document MN909, Pessoa marks the alternative directly as ‘Cancioneiro/Itinerario’ (see Figure 4).

Figure 5 shows a timeline of documents referencing the poem ‘Ode Marítima’ which was published in 1915. Nevertheless, the poem is mentioned in editorial lists that Pessoa wrote in later years, and under different titles, which suggests that the publication of a single work was not the decisive and final point in the planning and organization of his complete works.

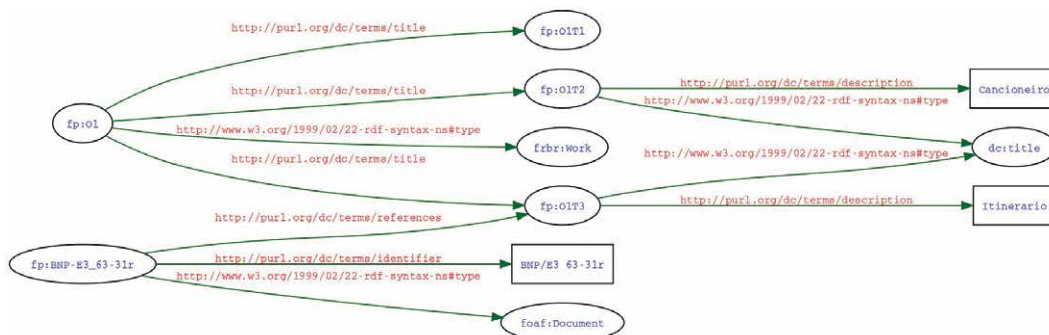


Figure 3: RDF graph showing relationships between a work, work titles and a document.

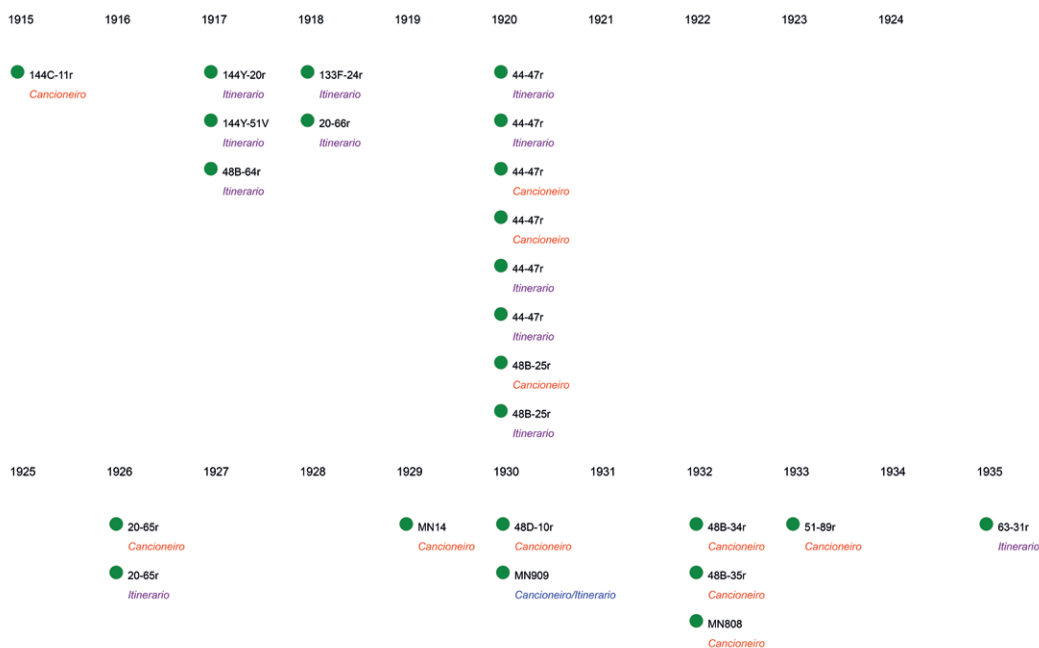


Figure 4: Timeline (1915-1935) showing document identifiers and alternative titles of the ‘Poems of Fernando Pessoa’ used in the documents.



Figure 5: Timeline (1915-1935) showing document identifiers and alternative titles of the poem 'Ode Marítima' used in the documents, as well as author and title of the published version.

Conclusion

The digital edition of Pessoa's editorial projects and publications draws on several forms of textual criticism stemming from different editorial approaches. Procedures used to create documentary, diplomatic, genetic and enriched editions are combined in order to make the most out of the material being edited. The TEI encoding of the source documents allows for the creation of coexistent forms of transcriptions as well as indexes and visualizations building on a simple genetic and semantic encoding. The combination of approaches primarily is determined by the research questions behind the edition: How did Fernando Pessoa plan and organize his complete works? How can his 'potential' work be analyzed? What conclusions can be drawn from his editorial projects regarding the question of the unitary or fragmentary character of his work? As has been shown, a digital edition lends itself particularly well to support the philological research concerned with the editorial projects and publications of Pessoa. The form of textual criticism proposed by this digital edition can be described as multiple: traditional and new forms of editing liaise to suit the subject matter of the edition and the characteristics of the edited material.

References

- Castro, Ivo. 2013. *Editar Pessoa*. 2nd. Edition. Lisbon: Imprensa Nacional-Casa da Moeda.
- Cunha, Teresa Sobral. 1987. 'Planos e projectos editoriais de Fernando Pessoa: uma velha questão'. *Revista da Biblioteca Nacional* 1 (2.2): 93-107.
- Deppman, Jed, Daniel Ferrer, Michael Groden (eds). 2004. *Genetic Criticism. Texts and avant-textes*. Philadelphia: University of Pennsylvania Press.
- Duarte, Luiz Fagundes. 1988. 'Texto acabado e texto virtual ou a cauda do come-
ta'. In *Revista da Biblioteca Nacional* 3 (2. 3): 167-181.
- Eide, Øyvind. 2015. 'Ontologies, Data Modeling, and TEI'. *Journal of the Text Encoding Initiative* 8. Accessed March 3, 2017. <http://jtei.revues.org/119>.
- Feijó, António M. 2015. *Uma admiração pastoril pelo diabo (Pessoa e Pascoaes)*. Lisbon: Imprensa Nacional-Casa da Moeda.
- Gabler, Hans Walter. 1999. 'Genetic Texts – Genetic Editions – Genetic Criticism or, Towards Discoursing the Genetics of Writing'. In *Problems of Editing, Beihefte zu editio 14*, edited by Christa Jahnson, 59-78. Berlin De Gruyter.
- Institut für Dokumentologie und Editorik (2014ff.), 'Questionnaire'. *Accompanying material to: ride. A review journal for digital editions and resources*. Accessed March 3, 2017. <http://ride.i-d-e.de/data/questionnaire/>.
- Lopes, Teresa Rita. 1992. 'A crítica da edição crítica'. *Revista Colóquio/Letras* 125/126, 199-218.
- Nemésio, Jorge. 1958. *A obra poética de Fernando Pessoa: estrutura de futuras edições*. Salvador da Bahia: Progresso Editora.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing. Theories, Models and Methods*. Farnham: Ashgate.
- Sepúlveda, Pedro. 2013. *Os livros de Fernando Pessoa*. Lisbon: Ática.
- Zenith, Richard (ed.) 2011. *Fernando Pessoa. Livro do Desassossego. Composto por Bernardo Soares*. Lisbon: Assírio & Alvim.

Reproducible editions

Alex Speed Kjeldsen¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Inspired by *Reproducible Research* I will argue for the production of what might be called *Reproducible Editions* and – more generally – for the usefulness of computing as much of an edition as possible, even outside a fully automated reproducible setup.

The main idea behind reproducible research is that data analyses, and more generally, scientific claims, are published together with their data and software code so that others may verify the findings and build upon them. In recent years this approach has gained popularity, not least in fields where statistics play an essential role.

When producing a scholarly edition, an article or an introduction to an edition in a reproducible way, we publish not only the text in its final format including the prose with possible figures and tables, but also the data (in our case typically annotated transcriptions) as well as the computer code used in the analytic work. This enables other users – including our future selves – to redo, build upon and adjust the work without the need to start over.

One of the great general advantages in a workflow based on reproducibility is how much easier one can verify the results, but in this paper I will focus on three more specific main points (also *cf.* *e.g.* Gandrud 2015, ch. 1).

First, a number of advantages are the direct consequence of analysis and writing being together. This in itself minimises the need for double bookkeeping, *e.g.* when analysing empirical material in an article or in the introduction to a scholarly edition. It is *e.g.* often the case that an error in the transcription or the linguistic or paleographical annotation is discovered only during the analysis, simply because you first notice how a certain feature sticks out when looking at it more systematically or analytically.

¹ alex@hum.ku.dk.

When discovering such errors or inconsistencies it is much easier to make changes in your interpretation of the empirical data when a reproducible workflow is applied, since changes in the data source will not require you to repeat the same more or less manual procedure. Furthermore you are not forced to use mental energy on the possible consequences of the individual changes for the perhaps 200 tables found in an introduction – something which is bound to go wrong anyway. Less manual work should generally result in less error-prone results, and it also opens up for the possibility to apply automatic tests to check the empirical data analysis (as known from software development).

Second, the adjustment of the workflow with the development of usable tools has some advantages in itself in the sense that you will actually develop better work habits. Even if you already write code to assist you in the analytic work, the demand to publish this code along with the empirical data pushes you to bring your data and source code up to a higher level of quality. At the same time it facilitates your own reuse of data and methods in new or revised investigations.

Third, a reproducible workflow opens up for better collaboration and use of the material, since other researchers are able to use the reproducible data or code to look at other, often unanticipated, questions. The fact that many things have to be stated explicitly in order for the reproducibility to work should furthermore facilitate better teamwork.

A number of tools exist which support reproducible research to varying degrees, but *org-mode* (orgmode.org) is probably the most versatile of these (*cf.* Schulte and Davison 2011; Schulte *et al.* 2011).

Org-mode is implemented as a part of the Emacs text editor, and initially it was developed as a simple outliner intended for note taking and brainstorming. Later it got support for – among other things – task management and the inclusion of tables, data blocks and active code blocks (*cf.* below).

Among *org-mode*'s strengths the following could be mentioned:

- 1) It is a purely text-based system with all the advantages this gives, *e.g.* in relation to version control and long term preservability.
- 2) It is an exceptionally great editing environment due to its integration with Emacs which has literally thousands of built-in commands for working with text. Among other things this facilitates very convenient editing of tabular data (incl. spreadsheet-like functionality), hierarchical structures such as lists and sections of different levels and a number of things with relevance for academic writing, *e.g.* functionality to insert and manage references and to create indices.
- 3) It has active code blocks, and *org-mode* is not only programming language agnostic in the sense that you can mix code blocks written in different languages in a single document, but also that code blocks can pass results and the contents of data structures to each other, effectively turning *org-mode* into a meta-programming platform.
- 4) It has great built-in, configurable export features with support for many formats and backends. *Org-mode* documents can therefore serve as a metaformat from which other formats can be generated.
- 5) It is extremely flexible since everything can be (re)programmed on the fly as a direct consequence of *Org-mode* being part of Emacs which has a Lisp interpreter as its core.
- 6) *Org-mode* is part of a very helpful and vivid open source community.

Because of this and the tight integration with Emacs, one of the oldest and most ported pieces of software in active development, it is likely to stay around and to be developed further for many years to come.

When using a reproducible workflow in which the computer extracts information from the transcriptions, it is especially advantageous to have quite specific markup. In my transcriptions of the oldest Icelandic original charters, for example, each individual word/token has the following kind of annotation: 1) Four levels of text representation (a facsimile level – a transcription very close to the actual forms in the charter, including abbreviations and various forms of special characters), a diplomatic level (abbreviations are expanded and some variant letter forms merged) and two normalised levels (Old and Modern Icelandic with fully standardised orthography). 2) Two lemmas (Old and Modern Icelandic). 3) Full morpho-syntactic annotation (part-of-speech). 4) Phono-graphemic annotation (mapping from the characters on facsimile level to the entities in a phonological reference system). 5) Paleographical annotation (of specific letter variants). 6) Linking of each grammatical word/token to the corresponding part of the digital image. 7) Non-linguistic annotation (e.g. persons and places, incl. geo-tagging). 8) Various kinds of notes.

Parts of the specific markup is illustrated in Figure 1. It is implemented in org-mode and encoded in a tabular format which makes it very easy to hide and show information as you see fit (by collapsing the individual columns and/or hiding individual rows), something which has proven invaluable when dealing with such extensive amount of meta-information for each word.

In a talk on reproducibility and computability it is important to stress that most of this markup has been automatically or semi-automatically generated. This is the case for the lemmas, the morpho-syntactic and grapho-phonemic annotation and three of the four levels of text representation.

All this information can be exported to TEI compatible XML following the guidelines from Medieval Nordic Text Archive (menota.org). It is also possible to export to other formats, *e.g.* html and pdf. You can export the entire corpus, specific charters or even more specific parts of the annotation, *e.g.* by producing word indices on the fly.

In my experience it is extremely useful to have the different kinds of functionality fully integrated into one system, not just in terms of practical reproducibility, but also because it facilitates interactive use in the editing process which is essential for the philologist who cannot automate everything to the same degree as the statistician.

u	eg	vér	xpe pi n ^o cM	vér	vér	W(er)	W	001	{W:v}{:}{ér,ér}
o	arni	Arni	xnp gM nS cN sI	Arni	Arni	arn(h)æ	arnæ	002	{a:[á,a]}{r:r}{n:n}{æ:I}
o	með	með	xap yb	með	með	w(ed)	m;	003	{m:m}{:}{eð}
o	guð	guð	xnc gM nS cG sI	guðs	guðs	guðz	guðz	004	{g:g}{u:u}{ð:ð}{z:s}
o	náð	náð	xnc gF nS cD sI	náð	náð	nadh	naðh	005	{n:n}{a:a}{ð:h}{ð}
o	biskup	byskup	xnc gM nS cN sI	biskup	byskup	byscoip	bp	006	{b:b}{:}{ysku}{p:p}
o	i	i	xap yb	i	i	i	j	007	{j:I}
o	skálholt	Skálholt	xnp gM nS cD sI	Skálholt	Skálholt	skalholti	fka(h)olti	008	{f:s}{k:k}{a:a}{l:l}{:}{h:h}{o:o}

Figure 1: Example of markup in Emacs' org-mode.

```

indikeres med en enkelt undtagelse (fyrær' 169.14) i; i alle 119 tilfælde med
/fyri(r)/ og sammensætninger hermed, fx fír' 211.10 og fírær biðu'
215.13 (for yderligere kommentarer til forholdene ved /fyri(r)/ og /yfi(r)/ med
perspektivering til forholdene i det øvrige diplommateriale, se afsnit
morf-prep).

I præf. af vb. /skulu/ er de to eksempler med {i} i første stavelse, fíldæ'
240.3 og fíldu' 274.17, registreret som tilfælde med {i} for {i} eller {y}. De
afspejler dog snarest i;. I de øvrige 10 eksempler med præf. af /skulu/
manifesteres rodvokalen som {u} x8, fx
fíldæ' 236.5 og
fíldæ' 223.10,
og {y} x2
fíldæ' 169.9 og fíldæ' 237.9.
Vokalen afhænger tilsyneladende ikke af modus (for yderligere kommentarer til
vokalforholdene ved /skulu/, herunder perspektivering til brugen i det øvrige
diplommateriale, se afsnit morf-verb-specifikke).
U(Unix)*- artikel.org 51% L7410 Git:master (Org Hi Helm Fill) to okt 6 06:33 0.12[+95%]
| fíldæ | fíldæ | 169-09 |
| fíldæ | fíldæ | 223-10 |
| fíldæ | fíldæ | 236-05 |
| fíldæ | fíldæ | 237-09 |
| fíldæ | fíldæ | 240-03 |
| fíldu | fíldu | 274-17 |

```

Figure 2: Interactive citation control.

```

*** (s) <<pal-s>> :DONE:...
*** (S) <<pal-ss>> :DONE:...
*** (t) <<pal-t>> :DONE:...
#+BEGIN_SRC emacs-lisp :exports none
(dipl-fon-generer-oversigt-tegn (point-min) (point-max) "t" "/home/ask/Dropbox/diplomer/afskrifter-JE-3.org")
#+END_SRC

##RESULTS:
##begin_example...
#+CAPTION: Brugen af (t)
#+attr_latex: :placement [h!]
#+LABEL: tab:pal-t
| (t) | t | 895 | | | (ðtt) | t | 1 |
| | b | 31 | | | (th) | t | 189 |
| | tt | 1 | | | | b | 1 |
| (ath) | á | 1 | | | (tt) | tt | 191 |
| (ðht) | t | 1 | | | | t | 95 |
| | tt | 1 | | | | t/tt | 1 |
| (ðt) | t | 5 | | | | | |
| (ðt) | t | 8 | | | | | |
| | d | 1 | | | | | |
| | tt | 1 | | | | | |

#+CAPTION: Græfklassem og -typer af (t)
#+attr_latex: :placement [h!]
#+LABEL: tab:pal-t-typer
| ||t|| | ||t|| | \includegraphics[scale=1.0]{/home/ask/Dropbox/diplomer/JE-billeder/164/164-008.png} | fíldæ | 164.1 |
| | | | \includegraphics[scale=1.0]{/home/ask/Dropbox/diplomer/JE-billeder/164/164-101.png} | fíldæ | 164.9 |
| | | | \includegraphics[scale=1]{/home/ask/Dropbox/diplomer/JE-billeder/239/239-188.png} | tok | 239.10 |
| ||t|| | \includegraphics[scale=0.9]{/home/ask/Dropbox/diplomer/JE-billeder/168/168-019.png} | gípuæt | 168.2 |
| ||t|| | \includegraphics[scale=1.5]{/home/ask/Dropbox/diplomer/JE-billeder/169/169-035.png} | þenofu | 169.2 |

Brugen af (t) er opsummeret i tabel tab:pal-t. Ud over de der anførte
tilfælde bruges (t) 59 gange i latinske ord. Endelig optræder det i 32 tilfælde
som del af refleksivformanten (normalortografiens /-sk/, /-zk/).
U(Unix)*- artikel.org 29% L4122 Git:master (Org Hi Helm Fill) to okt 6 06:43 0.15[+95%]

```

Figure 3: Table of grapho-phonemic relations generated by an active code block.

Interactive tools used for proofreading can serve as an example. When proofreading primary sources it is *e.g.* possible to take full advantage of the tagged image data. You can proofread more traditionally by automatically inserting images of the individual words into the table containing the transcription or perform more systematic proofreading on the basis of specific searches or queries. When proofreading an introduction to an edition you can check citations automatically, either for the whole introduction, for an individual citation or for all citations found in a specified part of the introduction. Also in this case, the actual words in the manuscript can be extracted from the digital images and viewed as part of the interactive citation control (*cf.* Figure 2).

The last example to be mentioned illustrates how elements of reproducibility can be used when producing an orthographic description. Instead of building up tables like the one seen in Figure 3 manually, they can be generated by embedding an active code block in the introduction. This code block is then responsible for extracting all the grapho-phonemic relations and for the layout of the table in its final form with caption, numbering etc. (*cf.* Figure 3). This means that changes in the transcription automatically will result in an updated table in the introduction.

The system that I have called MenotaB has been used on a daily basis for the last three years and will be developed further in the years to come. In regard to reproducibility the idea is to add functionality to generate the entire digital edition from a simple declarative description, including the html, css and JavaScript code.

This is especially fruitful because the edition will be self-contained with all functionality written in pure JavaScript (including all tools for advanced search, export and visualisation). Since it will not rely on external functionality, *e.g.* on a server, it makes dependency management, and therefore reproducibility and reuse, much easier.

An example of this kind of edition is my preliminary version of *Icelandic Original Charters Online* (IOCO). This edition has much of the same functionality built in as the MenotaB system, and often when new functionality is implemented, it is added both to the edition and to the editorial system.

If you want to use a reproducible approach it has some consequences for the philological workflow. Some important points are the following:

- 1) Use quite specific markup or annotation to facilitate better automatic data extraction.
- 2) Compute as much as possible (possibly combined with manual check of the computer's suggestions), both in terms of textual metadata and the accompanying prose.
- 3) Make the edition as self-contained as possible. Although not strictly necessary from a theoretical point of view, it makes things much easier in practice. With the powerful features of modern browsers with highly optimised JavaScript engines this is feasible.
- 4) Put everything under version control. This is particularly easy when using a text-based system with source code and data as part of the final product. You principally could have everything in a single org-mode document.
- 5) State the software dependencies very precisely (including the specific versions used). Ideally all software also should be packaged with the final product. Again, the more self-contained the product is, the easier this is to achieve.
- 6) The philologist does not need to become an expert programmer, but it is necessary to acquire some basic coding skills to be able to take full advantage of a programmable system.

Having asked what consequences it has for the philologist to use a reproducible workflow, we should also ask what consequences it has for the software we use. I think some of the most important demands for our software are the following: 1) It should be easy to compute (automatically). This means that we should generally not rely on the mouse or some other pointing device. 2) The software should be as adjustable as possible. It is essential that everything can be programmed, changed and enhanced with new functionality. 3) You should be able to interact easily with other tools and integrate them directly in the software and workflow. This includes tools for version control and the use of different programming languages. 4) All software should be free and open source. 5) It should be easy to package things. As already mentioned, this most easily can be achieved by using systems which are as self-contained as possible. This on the other hand stresses the importance of the high degree of adjustability, since it is impossible (and probably not even desirable) to predict all future needs.

References

- Gandrud, Christopher. 2015. *Reproducible Research with R and R Studio. Second Edition*. Chapman and Hall/CRC.
- Schulte, Erik and Dan Davison. 2011. 'Active Documents with Org-Mode'. *Computing in Science & Engineering* 13.3: 2-9.
- Schulte, Erik, Dan Davison, Thomas Dye and Carsten Dominik. 2012. 'A Multi-Language Computing Environment for Literate Programming and Reproducible Research'. *Journal of Statistical Software* 46.3: 1-24.

'... but what should I put in a digital apparatus?' A not-so-obvious choice

New types of digital scholarly editions

Raffaella Afferni,¹ Alice Borgna,² Maurizio Lana,³

Paolo Monella⁴ & Timothy Tambassi⁵

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

We propose to develop and expand the concept of 'digital edition of a text'. The specific value of a digital edition is not only in the digital form of representation of textual information: dynamic rather than static, resulting in better visual or practical usability, but it mainly lies in the ability to work with computational methods on the text and on the information it conveys. Therefore the digital edition of a text should aim to provide adequate data and functionality to further forms of processing.

Hence the idea that the 'digital scholarly edition' until now often identified with the 'digital critical edition' (i.e. an edition *variorum*, reporting variant reading), also can take other forms focused on other types of 'scholarly research': from the geographical knowledge contained in the text, to the historical knowledge (time and events) often inextricably linked with the prosopography, and much more.

If the *digital critical edition* is a type of *digital scholarly edition* containing an apparatus that analyses and describes the state of the text in the witnesses, then we can conceive e.g.

1 raffaella.afferni@uniupo.it.

2 alice.borgna@uniupo.it.

3 maurizio.lana@uniupo.it.

4 paolo.monella@gmx.net.

5 timothy.tambassi@uniupo.it.

- the *digital scholarly geographical edition* of a work – whose apparatus contains an analytical description of the geographical knowledge contained in the placenames;
- the *digital critical geographical edition* whose geographical apparatus is layered over a base critical edition:

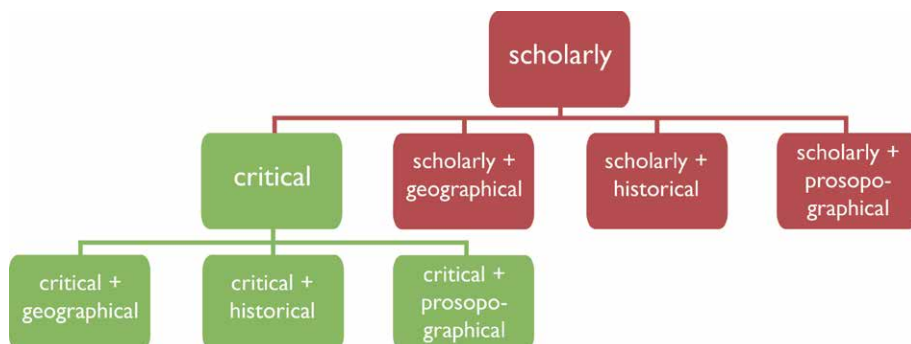


Figure 1: Types of digital scholarly editions.

By ‘base critical edition’ we do not necessarily mean a previously published edition with a critical text already established. The best model would be an integrated edition where the critical discussion on (and selection of) textual variants and the interpretive extraction of geographical knowledge are integrated and both ‘born digital’.

The knowledge contained in the text must be expressed in a highly formal manner – the same way that the critical apparatus is a highly formal device – by means of an ontology. The ontology both from a philosophical or a computer science point of view is a structure aimed to analyse and describe the categorical hierarchy of a specific domain, analysing its basic constituents (entities like objects, events, processes, etc.), the properties characterizing them and the relationships which correlate them. The resulting (structural) representation of knowledge allows to resolve conceptual or terminological inconsistencies, providing a dictionary of terms formulated in a canonical syntax and with commonly accepted definitions. It also provides a lexical or taxonomic framework for the representation of knowledge, shared by different communities of information systems that can range across several domains.

From this point of view, the starting point can be the adoption of GO!, a geographical ontology aimed at providing a complete and informative description of the geographical knowledge emerging from Latin literature.⁶ The most general aims of GO! are essentially three: accessibility (both for the scientific community and for general public), informativeness and completeness. Moreover, about the most specific goals, GO! has been developed to describe the geographical locations, with a particular attention to the description of the Ancient World, especially to give the opportunity of having a link between the places mentioned

6 <https://goo.gl/3VRPGt>.

in the texts, especially ancient, and their identification and correspondence with contemporary ones. For classical scholars this correspondence of ancient and contemporary modelling is of undisputed interest, both for the study of the habits of the most ancient peoples, and for the most various themes of literary interest. Through ontologies you can build maps of the ancient world and compare them to contemporary ones, annotate historical, geographical, cultural details connected to the place, indicate in which ancient text the place is mentioned and which author discloses the details. These are just some ideas for research that can be developed, but the scenario that opens through these connections will be much larger.

From a scholarly point of view we also can add that digital critical editions of classical works whose textual tradition is made of many witnesses are still very rare. The ancient literatures scholars usually ask to the digital no more than authoritative collections of texts (TLG, PHI, and online digital libraries). So the opportunity to enrich the digital text with variants (especially from a new collation of manuscripts) has known little practical application. Even less common in Classics, not to say absent, is the model of an edition 'based on full-text transcription of original texts into electronic form' (Robinson 2006). The peculiar nature of textual variance in classical texts, where the discarded lesson is a mistake to recognize and remove, contributes to this closure face to the opportunities of the digital. Consequently a digital critical edition aimed to include a bigger number of variants – that is 'errors' – than in printed format is unsustainable in terms of cost/benefit evaluation. Thus a new space for reflection opens, no longer confined to the form (that is to the textual tradition) but open to the content of the text formally analysed in the apparatus, which might be thought of as a space open to contain other, new, kinds of knowledge.

References

- Buzzetti, Dino. 2002. 'Digital Representation and the Text Model'. *New Literary History* 33 (1): 61-87.
- . 2009. 'Digital Editions and Text Processing'. In *Text Editing, Print, and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland. London: Ashgate, 45-62.
- Monella, Paolo. 2012. 'Why are there no comprehensively digital scholarly editions of classical texts?' Paper presented at the *IV Meeting of digital philology* (September 13-15, Verona). Accessed March 3, 2017. http://www1.unipa.it/paolo.monella/lincei/files/why/why_paper.pdf.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing. Theories, Models and Methods*. London: Ashgate.
- Robinson, Peter. 2006. 'The Canterbury Tales and other Medieval Texts'. In *Electronic Textual Editing*, edited by Lou Burnard, Katherine O'Brien O'Keefe, John Unsworth. New York: The Modern Language Association of America. 128-155. Online: http://www.tei-c.org/About/Archive_new/ETE/Preview/.

Critical editions and the digital medium

Caroline Macé¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Many critical editions are now digital, especially those which are no longer under copyright: as PDF or in large digital text corpora (normally without apparatus and philological introduction). New critical editions, born digital, are rarer. Scholarly editors of ancient and medieval texts are reluctant to publish their editions online, because so far there is no convincing replacement for the role that publishing houses used to play: giving credit, visibility, recognition to the published work and some confidence that it may last and be found in libraries. Compared to the well-organized world of published scholarship, with its book series, reviews, repertories, encyclopaedias, the rather chaotic and anarchic world of internet publishing is still unappealing and adventurous. In addition, the current model of digital scholarly editing, at least the one that is attracting most attention, seems mostly suitable for documentary editions, which is only one type of edition fitting one type of textual tradition. Of course, it is wonderful to have digital images facing diplomatic transcriptions (or transliterations) of documents, especially when they preserve rare or unique texts. This had been done long ago, although at that time on paper, by papyrologists, or people working on palimpsests, and even digitally as early as 1997 in the case of the edition of Tocharian fragments in the text corpus 'Titus' (<http://titus.fkidg1.uni-frankfurt.de/texte/etcs/toch/tocha/tocha.htm>). But not all ancient or medieval texts require the same treatment and we cannot replace critical editions by a juxtaposition of diplomatic editions somehow 'automatically' (?) compared (?), as some people seem to believe.

Those of us who have been working with black-and-white microfilms of manuscripts, sometimes barely readable, especially with those poor microfilm-readers we had then, know how great it is to be able to work with high-quality

1 mace@em.uni-frankfurt.de.

digital images of manuscripts. Many manuscripts still can be viewed only thanks to microfilms (sometimes very old ones), however, and microfilm-readers are becoming increasingly rare in libraries and universities (and usually no one knows how to use them). Even digital reproductions are not always faultless, not to speak about the fact that it is sometimes difficult to know that they exist in the first place and under which name you can call them on the website of the library. Sometimes the rectos and the versos have been put together in the wrong order (since manuscripts often have the bad habit of being foliated), or one page was not photographed, when it is not the old bad microfilm which has been digitised instead of the manuscript itself... At any event, and with all those reservations, no one can deny that the digitization of manuscripts represents a considerable help for scholarly work – but in itself this is not a scholarly achievement, but rather a technical one, carried out thanks to institutional well-willingness.

Nobody will deny either that producing an accurate transcription or transliteration of a text preserved in a document is by itself a scholarly work, especially when that document is difficult to read. Just to take one extreme example, imagine how difficult it is to read a text written in a language that nobody knows using a script that nobody knows, and that as an under-text that has been washed away and overwritten (Gippert *et al.* 2009). But whereas one can admire the technical performance of aligning images and transcriptions of, for example, in the case of the famous ‘Codex Sinaiticus’, the Greek Bible (<http://www.codexsinaiticus.org/de/>), it must be said that already in 1862 (Tischendorf) a transcription of the ‘Codex Sinaiticus’ existed, that has been corrected and improved by many scholars since then, and even a facsimile was printed in 1911 and 1922 (Lake). So, is it wicked not to be impressed by the novelty of that scholarly endeavour (especially since the reference system of the Bible has been standardized thanks to centuries of scholarship)? Many transcriptions of manuscripts made by outstanding scholars in the 19th or early 20th century (and before) exist in published form (for example Cramer 1835-1837 and 1839-1841). They were made by scholars who had the chance to work on manuscripts and wanted their colleagues to be able to read the texts without having to undertake an expensive trip to the libraries where those manuscripts were kept. They transcribed texts in which they were interested, which seemed unusual or rare. Often people do not know that they exist, unless they have been noted in repertories, because they are not really editions and therefore are not published in the same way. It would be great to have in digital form not only images of manuscripts, but also as many transcriptions as possible, not only new ones, but also old ones, and have them in a place and with a reference system that makes it easy to find them and to refer to them.

However, with or without those digital images and digital transcriptions, the scholarly work of engaging critically with the text has to be done, and its results have to be expressed in a suitable way, in order to be in turn read, understood and criticised by other scholars. Digital tools may help as well in that endeavour, although in a rather limited way, but they are still few and often not accessible to scholars who are not themselves programmers (Andrews and Macé 2015). And so far nothing has replaced the critical apparatus (although it may appear as a pop-up

box on the webpage rather than at the bottom of the printed page) as the most economic way of representing textual variation.

There is still a long way to go before digital critical editions will be more appealing to scholars producing them (Macé and Gippert forthcoming).

References

- Andrews, Tara L. and Caroline Macé (eds). 2015. *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols.
- Cramer, Johan Anthony. 1835-1837. *Anecdota Graeca e codicibus manuscriptis bibliothecarum Oxoniensium descripta*, 4 vols. Oxford.
- Gippert, Jost, Wolfgang Schulze, Zaza Aleksidze and Jean-Pierre Mahé. 2009. *The Caucasian Albanian Palimpsests of Mount Sinai*, 2 vols. Turnhout: Brepols.
- Lake, Kirsopp. *Codex Petropolitanus Sinaiticus et Friderico-Augustanus Lipsiensis*, 2 vols., 1911 & 1922. Oxford.
- Macé, Caroline and Jost Gippert. Forthcoming. Textual Criticism and Editing in the Digital Age. In *Oxford Handbook of Greek & Latin Textual Criticism*, edited by Wolfgang de Melo & Scott Scullion. Oxford: Oxford University Press.
- Tischendorf, Constantin. 1862. *Bibliorum Codex Sinaiticus Petropolitanus*, 4 vols., Petersburg.

Scholarly editions of three rabbinic texts – one critical and two digital

Chaim Milikowsky¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Introduction

One of the main foci of my scholarly research for the past thirty or so years has been textual criticism in a very broad sense, studying such matters as transmission and tradition, identifying texts and fragments, assessing and establishing textual traditions, studying the work of copyists as well as textual contamination and corruption, and – in the final instance – modes of editing a text.

I have been involved in various capacities in the production of three scholarly editions of rabbinic texts, and will give a short outline here of the nature of the works and the modes of textual presentation used in each of the three editions.

But first I wish to explicate my title 'Scholarly editions of three rabbinic texts – one critical and two digital', and define clearly the underlying distinction between a scholarly edition and a critical edition. A scholarly edition is any edition which presents in some form or another the manuscript material of the work under study. This can be a diplomatic edition of the text of one document of a work, or a variorum edition, presenting the texts of a multitude of documents of a specific work in a (hopefully) easy-to-use format. A critical edition, however, presupposes the use of the critical faculty of the human mind to reconstruct – or perhaps it is more fitting to write, construct – a better text of the work than any that has been preserved in the extant documents. Any critical edition is, by definition, a scholarly edition; many scholarly editions are not critical editions. (It should be clear by now that my remarks in this short essay have little or no relevance to works created in the post-print era).

¹ chaim.milikowsky@biu.ac.il.

Note the three terms I am using here: work, document and text. The term work is being used to denote the author's or editor's product (one which theoretically may never have existed in any concrete mode of expression such as a manuscript or book). A document is a concrete mode of expressing a work: thus the manuscripts of the *Illiad* are the documents of the work conceptualized/reified as the *Illiad*. The text of a work is the actual word-after-word presentation of the original product and the text of a document is the word-after-word presentation found in a manuscript. Generally, texts of documents are used to try to reconstruct texts of works, although it is of course legitimate, and very often important, to take as one's goal the presentation of the text of a specific document. For the purposes of this discussion I am ignoring the theoretical position that there is no such thing as the text of a work or that it is invalid to try to reconstruct the texts of works and one can deal only with the texts of documents. Against this position I can do no better than refer the reader to the still useful and very admirable short book by G. T. Tanselle, *A Rationale of Textual Criticism* (Tanselle 1989). Or to quote the classicist Martin West who wrote in his *Studies in the Text and Transmission of the Iliad*: 'As regards the text, I conceive the aim as being the best approximation that may be possible to the *Iliad* as its original author left it' (West 2001, 158).

My crucial distinction then is between non-critical scholarly editions and critical editions. Regarding non-critical scholarly editions I think it is immediately obvious to everyone that these days there is no justification for the print publication of such editions, whether the edition is a diplomatic transcription of the text of a specific document or it presents the texts of many documents together in one format or another. The digital mode of presentation takes little space – the bane of diplomatic transcriptions in general – and also allows easy access and easy revision, in a sense democratizing admittance to what used to be a small coterie of scholars who were at elite institutions with extensive library holdings in all scholarly fields.

I am less sure that critical editions should be produced in digital format. I will return to this point below.

And now to the three scholarly editions.

Scholarly edition no. 1: A critical edition of Seder Olam (Chaim Milikowsky, *Seder Olam: A Critical Edition with Introduction and Commentary*, 2 volumes, Jerusalem: Yad Ben Tzvi 2013).

My work on this edition began as part of my doctoral dissertation, submitted to Yale University in 1981.

Seder Olam is an early rabbinic chronography of the biblical period whose primary *raison d'être* is dating events not dated in Bible, from the earliest parts of the biblical narrative (e.g., the Genesis accounts of the Tower of Babel and the Binding of Isaac) to the latest (e.g., the year of Jehoiaquim's reign in which Nebuchadnezzar ascended to the throne), and resolving chronological cruces, such as synchronizing the regnal lists of the kingdoms of Judah and Israel and resolving contradictions between the biblical books of Kings and Chronicles. It is the earliest preserved rabbinic composition and was composed in the 1st-2nd centuries of the

מאדם ועד המבול אלף ושש מאות וחמשים ושש שנה. חנוך קבר אדם והיה אחריו חמשים ושבע שנה. מתושלח מיצה ימיו עד המבול. מן המבול ועד הפלגה שלוש מאות וארבעים שנה. נמצא נח היה אחר הפלגה עשר שנים. אבינו אברהם היה בפלגה בן ארבעים ושמונה שני. א' ר' יוסה נביא גדול היה עבר שקרא את שם בנו פלג ברוח הקדש שני כי בימיו נפלגה הארץ (בר יוכה) ואם תומר בתחילת ימיו הלוא יקטן אחי (יו) היה קטון ממנו והוליד משפחות וניתפלגו ואם תומר באמצע ימיו והלא לא בא הכתוב לסתום אלא לפרש הא אינו או' כי בימיו נפלגה הארץ אלא בסוף ימיו.

אברהם אבינו היה בשעה שנדבר עמו בין הבתרים בן ע' שנה שני ויהי מקץ שלשי(ם) שנה וארבע מאות שנה וגו' (שמי יב: מא) והזר לחרון ועשה שם חמש שני ואברהם בן חמש שני(ים) וש(בעים) שנה) בצ(אחו) מח(רן) (בר יב: ד). נמצא מן הפלגה) ועד שיצא אבר(הם) מח(רן) עשרים ושש ש' (שנה) בצ(אחו) מח(רן) (בר יב: ד).

3 מן ... שנה². בראשית רבה כו: ג (עמ' 246); משנת רבי אליעזר, ח (עמ' 145). 4-8 א' ... ימיו. בראשית רבה לז: ז (עמ' 349). עיין להלן כא: 25. 9-11 אברהם ... מח(רן)¹. עיין במדבר רבה יד: יא; משנת רבי אליעזר, ב. מדה לג (עמ' 40); המקרא בבית הכנסת, ב. עמ' קעו.

1 *פרק א] יצא במ סוד עולם נ סדר עולם ד הק סדר עולם פרק ראשון ו אתחיל סדר עולם בסדר ו אתחיל סדר עולם בעזרת נסתר ונעלם א תניא דסדר עולם ל ספר סדר עולם פ 2 המבול נח קנח אלף וחמשים ושש שנים מנח ועד שם חמש מאות שנים משם ועד ארפכשד מאה שנה אחר המבול שנתים נולד ארפכשד מאדם ועד המבול פ ו חמשים ושש] יצא ק | שנה] יצא ב שנים א שנים זה (זהו ל) פרטן אדם ק"ל שת ק"ה אנוש צ' קינן ע' מהללאל (מהלל ל) ס"ה ירר קס"ב חנוך ס"ה מתושלח קפ"ז למך קפ"ב (פ"ב ל) נח בן שש מאות שנה וגו' דל | קבר | קבר את אבד הלק | אדם | אדם וחוה מ פ | וזיהו וחי ל | חמשים ושבע] יצא ק 3 שנה¹ שנים אב | *מיצה ימיו דל | היה נ מוציא שנותיו מקר מוציא משנותיו ומשנותיו ה הוציא שנותיו אב | מן | ומן ק | ועד | עד בד ל ועד דור | | שלוש מאות וארבעים] ש' ק 4 חיה חי לק | אחר | אחרי בל | אבינו אברהם | אברהם אבינו ב הלק אברהם א | בן | בן ל | שני] שנה ד הלק שנים א | יוסה | יוסי כל שאר כתבי היד 5 עבר עובר ר | את | יצא אבד לר | שם בנן לבנו ב | כי ... הארץ] שם הארץ פלג כי ... הארץ א כי ... הארץ וגו' ד 6 ואם תומר¹ ואם תאמר ה ל אם תאמר אקר אם ד וא"ת והלא ב | ימיו

ימיו היה קר | הלוא] יצא ב והלא אד הלקר | יקטן | קין ק | היה] יצא ה ל | קטון | קטן אבד הלק | ממנו | ממנו היה | משפחות | י"ג משפחות דל כמה משפחות ה משפחות הרבה א | *וניתפלגון וניתפלגן נ | תומר² | יצא ד תאמר אב הלקר 7 באמצע] יצא ל | ימיו] יצא ק ימיו היה הר | והלא ל | יצא אב הלקר הא מ | *הכתוב אבה נקר | יצא דל | לסתום] יצא ל | אלא לפרש] יצא ר | הא ... או' לנ | יצא אהר הא לא אמ' בק | כי ... הארץ אב לנ | יצא המפקר | אלא² | יצא א 8 ימיו היה ה 9 אברהם אבינו | אבינו אברהם אה | שנדבר הלקר | שנדברו נ שרבר אבמפ | עמו הן עמו הקב"ה א הקב"ה עמו ב עמו שכינה ל לו המקום ב' ה' ק | בין ... ע' | בין שבעים | שני] יצא ר | שנה² | יצא ק 9-10 וארבע מאות שנה] יצא ל 10 וגו' | יצא אבהר | וחזר חזר אבהר לאחר שנדבר עמו ירד דל | שם] יצא ל | חמש¹ | יצא ב | שני] יצא ר שנים שני בד הלק שנים שכן הוא או' א | בן | בין ל | שני(ים) | יצא ב 11 נמצא נמצאת או' א | מן הפלגה) | מהפלגה ה | ועד | עד בהקר | אבר(הם) | אברהם אבינו אד הקר אבינו אברהם ל | עשרים ושש אבה פ כ"ז דל עשרים ושנים ר ל"ז ק

Figure 1: A standard critical edition in print format: Seder Olam.

Common Era; later works cite it thousands of times. No critical edition of the entire work has been published ever.

The text of the edition was established using classic stemmatological methods. A codex optimus was presented as the base text, but was changed freely when the stemmatic evidence was definitive. It is based upon all extant manuscripts and early printed editions. Variants are presented in classical critical edition mode (see Figure 1). No digital critical edition has been prepared. (A digital image of a paper critical edition cannot be called a digital critical edition). A primary reason for this lack is the simple fact that this edition is, as noted above, an outgrowth and extensive revision of an edition produced well before the digital age began, but this is not the only reason.

Scholarly edition no. 2: A synoptic line-under-line edition of *Vayyiqra Rabba*

Vayyiqra Rabba is a homiletic-cum-exegetic early midrashic work with Leviticus as its base, and is a central repository of rabbinic ideology and theology. It, like most classical rabbinic works, is not the work of an author, but a collection of homiletical sermons, exegetical explications, short narratives, pithy apothegms, and other similar genres collected together, compiled and variously revised by an editor sometime in the 5th century of the Common Era.

Given the compilatory nature of rabbinic literature, and given the fact that this literature is replete with parallel passages appearing in two or more works, from both practical and theoretical vantage points it becomes difficult to determine if it is legitimate to speak of a redactorial moment with regard to these works, or should one speak more of works in constant states of becoming/changing.

My own research has led me to conclude that at least with regard to Vayyiqra Rabba (and a number of other midrashic works whose textual traditions I have studied in more or less detail), it is eminently correct to speak of a redactorial moment.

A critical edition of Vayyiqra Rabba was published by Mordechai Margulies (1972). Nonetheless, the difficulties in analyzing a not-simple textual tradition using solely the variant readings cited in a critical apparatus are well-known, and for my work on Vayyiqra Rabba, I found it essential to produce a synoptic edition, with the variant texts presented line under line (Figure 2). The beginnings of this edition were in a project jointly directed by the late lamented Professor Margarete Schlüter (of the Goethe University of Frankfurt) and me, and funded by the German-Israel Foundation.

לונ	{א}	ר' תנחום בר' חנילאי
מינ	{א}	ר' תנחום בר' חנילאי
פריז	{א}	ר' תנחום בר' חנילאי
דפוס	{א}	ר' תנחום בר' חנילאי
ג5	{א}	<...>
ירו1	{א}	ר' תנחום בר' חנילאי
או3	{א}	ר' תנחום בן חנילאי
או51	{א}	ר' תנחום בן חנילאי
ששון	{א}	ר' תנחום בר' חנילאי
קפ	{א}	ר' תנחום בר' חנילאי
<hr/>		
לונ	פתח	ברכו יי מלאכיו גיבורי כח עושה דברו
מינ	פתח	ברכו יי מלאכיו גבורי כח עושה דברו
פריז	פתח	ברכו יי מלאכיו גבורי כח עושי דברו
דפוס	פתח	ברכו יי מלאכיו גיבורי כח עשי דברו
ג5	פתח	ברכו יי <...> גיבורים <...> עושי
ירו1	פתח	ברכו יי מלאכיו גבורי כח עושי דברו
או3	פתח	ברכו יי מלאכיו גבורי כח עושי דברו
או51	פתח	ברכו יי מלאכיו גבורי כח עושי דברו
ששון	פתח	ברכו יי מלאכיו גבורי כח עושי דברו
קפ	פתח כתי'	ברכו יי מלאכיו גיבורי כח עושי דברו
<hr/>		
לונ	לשמ' קול דברו	במה הכת' מדבר אם בעליונים
מינ	וגו'	במה הכת' מדבר אם בעליונים
פריז	לשמע בקול דברו	במה הכת' מדבר אם בעליונים
דפוס	וכו'	במה הכתוב מדבר אם בעליונים
ג5	אמ' ר' <...>	אם בעליונים
ירו1		במה הכת' מדבר אם בתחתונים'
או3		במה הכתוב מדבר אם בתחתונים
או51		במה הכתוב מדבר אם בתחתונים
ששון		במה הכת' מדבר אם בתחתונים'
קפ		במה הכת' מדבר אי בתחתונים
<hr/>		
לונ	הכת' מדבר והלא	כבר נא' ברכו יי כל צבאיו אם בתחתונים
מינ	הכת' מדבר והלא	כבר נאמר ברכו יי כל צבאיו אם בתחתונים
פריז	הכת' מדבר והלא	כבר נאמ' ברכו יי כל צבאיו
דפוס	הכתוב מדבר והלא	כבר נאמ' ברכו יי כל צבאיו ואם בתחתונים
ג5	הכת' מדבר והלא	כבר נאמ' ברכו <...> בתחתונים
ירו1	הרי	כבר נאמ' ברכו יי כל מעשיו אם בעליוני'
או3	הרי	כבר נאמר ברכו יי כל מעשיו אם בעליונים
או51	הרי	כבר נאמר ברכו יי כל מעשיו אם בעליונים
ששון	הרי	כבר נאמ' ברכו יי כל מעשיו אם בעליוני'
קפ	הרי	כבר נאמ' ברכו יי כל מעשיו אם בעליונים

Figure 2: An online line-under-line synoptic edition: Vayyiqra Rabba.

The digital edition was input using plain-text (Hebrew) ASCII; care was taken to add no codes or coding (forestalling incipient obsolescence).

Inasmuch as this edition does not purport to reconstruct any 'original text' or a hypothetical archetype/hyperarchetype it is a classic instance of a scholarly edition, published online for easy access to all.

Scholarly edition no. 3: The Friedberg Project for Babylonian Talmud Variants ('Hachi Garsinan') (commenced c. 2012)

The project will encompass all Babylonian Talmud textual witnesses, i.e. manuscripts, early printings, Genizah fragments, binding fragments and other fragments found in public libraries and private collections all over the world and will include all tractates of the Babylonian Talmud. It presents high quality digital images of all original text-witnesses, accompanied by precise transcriptions of the text in the image. It will display the text-witnesses by column synopsis as well as by row synopsis, dynamically, enabling the user to choose which variants to highlight and which to omit, while emphasizing the differences between the text-witnesses, using advanced visual methods that will help the user complete his quest quickly and efficiently. It also will integrate additional functions including full text search on all text-witnesses, as well as save, copy and print options, personal workspace,

The screenshot shows a web interface for a Babylonian Talmud variant database. At the top, there's a navigation bar with 'ברכות < דף ב' and a search icon. Below this, a header row identifies the columns: 'דפוס וילנא' (Vilna edition), 'דפוס ונציה 1520 - 1523' (Venice edition), 'דפוס שונוצין 1484 - 1489' (Shonits edition), and '...T-S F 1(2).1'. The main text area displays a columnar synopsis of text variants in Hebrew. The text is arranged in four columns, each representing a different textual witness. The interface includes a sidebar on the left with navigation tools like 'סינופטי', 'בטורים', 'סינופטי', 'בשורות', 'סינופטי', 'במש', 'סינופטי', 'ללמד', 'ערכת יחידו', 'סינופטי', 'הצג קבוצה', and '123'. The bottom of the interface shows a 'גמ' (Gemara) section with a 'גמ' button.

Figure 3: A web database with a columnar and a linear synoptic edition and interactive access to digital photos and transcriptions: Babylonian Talmud.

etc. A partial version began operation approximately a year ago (March 2015) and is freely available, though registration is required (Figure 3).

This is an audacious project, from many perspectives, and could not have been embarked on without very generous private funding. Just locating all the witnesses to the Babylonian Talmud, receiving permission to use them and assembling high-resolution digital images of all these manuscripts and early prints has necessitated large outlays of money and time. And then began the crucial work of transcribing all these images, until the staff and the Academic Advisory Board were confident in the accuracy of the digital text. In parallel, a computing team began work on the various software and network aspects of the project.

This is also an edition which has as its goal the presentation of the source material and has no pretensions of creating a critical text.

Points to Ponder in Conclusion

There is, as indicated, a world of difference between the first scholarly edition noted above, on the one hand, and the second and third, on the other. The first is a critical edition, and presents a text which to my mind is as close to the text of the original work as one can get using the extant documents. The other two editions are scholarly non-critical editions, whose sole goal is to present to as wide an audience as possible the texts contained in the documents of the work. Digital scholarly non-critical editions have many advantages. As well as saving trees and bookshelf space, they have the advantage of easily being updated.

A printed edition cannot be changed; if a new manuscript is discovered, the only way to incorporate it is to produce a new edition. With regard to a digital edition, the solution is relatively trivial.

It is fitting that the first edition is paper-bound and the others digital. An edition printed on the correct materials has the capability to last centuries, but will digital editions last centuries?

Crucial questions relating to preservation, sustainability, and long-term access, all closely inter-related, have been raised again and again by those involved with the production of digital editions, and the answers we have are sadly deficient.

Let me give some depressing examples. We are burdened with a growing number of scholarly editions, digital castaways from as recently as the 1990s. The CD-ROM multiple edition of Johnson's Dictionary, for example, does not work with the operating systems of many post-2005 computers, and will become increasingly unusable as those systems continue to evolve. The Samuel Hartlib Papers, offered in digital format on two CD-ROMs in 1996 (edited by M. Greengrass and M. Hannon) was overtaken quickly by advances in software, reprogrammed, and again rendered obsolete by later updates. It languishes in the digital doldrums. In the UK, funding organizations such as the Arts and Humanities Research Council have become reluctant to support applications for mounting primary materials on line in recognition of the fact that these high-cost projects simply disappear without frequent, and often very expensive upgrades.

Nor is it likely that we will have the hardware to display today's digital images on a screen in 50 years, let alone 500. Remember floppy disks? And will the

information encoded on chunks of silicon, including hard drives, be electronically legible 50 years from now? The odds are, no. Ordinary CDs, DVDs, and hard drives have shelf-lives of 7-10 years. They are vulnerable to oxygen and other elements in the air, to sunlight, dust, and magnetism. The most optimistic manufacturers of expensive, solid-gold coated CD-ROM's express hope that their products may last for 100, even 200 years! At best, that's less than half the shelf life of acid-free paper.

Digital images of anything, left to their own devices, as it were, turn out to be a lot more fragile, a lot less durable over time, than papyrus or paper, let alone parchment. Digital is still the technology of choice, but it is unfortunately the most ephemeral means of textual transmission since man wrote on sand. Instead of recopying everything important every few centuries, digital editions must be recopied and their software revised at the minimum every few decades, to insure readability. And that readability will still depend on access to appropriate hardware.

An important figure in the field of modern literary studies who has placed digital editing, processing and study at the center of his scholarly career is Jerome McGann, author of the classic and controversial *A Critique of Modern Textual Criticism* (McGann 1983). In what I see as a rather sad epilogue to much of what he has done (he was born in 1937), he recently wrote:

I spent eighteen years designing The Rossetti Archive and filling out its content. This was a collaborative project involving some forty graduate students plus a dozen or more skilled technical experts, not to speak of the cooperation of funding agencies and scores of persons around the world in many libraries, museums, and other depositories. It comprises some 70,000 digital files and 42,000 hyperlinks organizing a critical space for the study of Rossetti's complete poetry, prose, pictures, and designs in their immediate historical context. The Archive has high-resolution digital images of every known manuscript, proof, and print publication of his textual works, and every known or accessible painting, drawing, or art object he designed. It also has a substantial body of contextual materials that are related in important ways to Rossetti's work. All of this is imbedded in a robust environment of editorial and critical commentary. (...)

On the other hand, if the Archive is judged strictly as a scholarly edition, the jury is still out. One simple and deplorable reason explains why: no one knows how it or projects like it will be or could be sustained. And here is the supreme irony of this adventure: I am now thinking that, to preserve what I have come to see as the permanent core of its scholarly materials, I shall have to print it out.

(McGann 2010; emphasis added).

It should be noted immediately that in spite of these apocalyptic reflections, digital editions still are being funded, still are being produced. I wish their progenitors well, and hope my Cassandra-like musings are all mistaken.

References

- Greengrass, Mark, and Michael Hannon (eds). 1996. *The Samuel Hartlib Papers*. CD-ROM. Michigan: University Microfilms.
- Margulies, Mordechai, (ed.) 1972. *Vayyiqra Rabba*, 3 vols. 2nd printing, Jerusalem: Warhmann Books.
- McGann, Jerome. 1983. *A Critique of Modern Textual Criticism*, Chicago: University of Chicago Press.
- . 2010. *Sustainability: The Elephant in the Room*. <http://cnx.org/contents/PVdH0-ID@1.3:AeBKUQq-@2/Sustainability-The-Elephant-in>. Accessed on 5 June 2016
- Milikowsky, Chaim, and Margarete Schlüter (eds). 2005. *Synoptic edition of Vayyiqra Rabba* <http://www.biu.ac.il/JS/midrash/VR/>. Accessed on 4 March 2017.
- Milikowsky, Chaim. 2006. 'Reflections on the Practice of Textual Criticism in the Study of
- Midrash Aggada: The Legitimacy, the Indispensability and the Feasibility of Recovering and Presenting the (Most) Original Text. In *Current Trends in the Study of Midrash*, edited by C. Bakhos. 79-109. Leiden: Brill.
- . 2013. *Seder Olam: Critical Edition, Commentary and Introduction*, 2 vols. Jerusalem: Yad Ben Tzvi.
- Tanselle, George Thomas. 1989. *A Rationale of Textual Criticism*. Philadelphia: University of Pennsylvania Press.
- West, Martin. 2001. *Studies in the Text and Transmission of the Iliad*. Munich: De Gruyter.

From manuscript to digital edition

The challenges of editing early English alchemical texts

*Sara Norja*¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

Alchemy, later considered a pseudo-science, was one of the first experimental sciences in the Middle Ages and influenced the development of chemistry. A multitude of English medieval alchemical manuscript texts survive, written in both Latin and the vernacular. However, the uncharted material vastly outnumbers the texts edited so far, especially in the case of vernacular texts. Indeed, according to Peter J. Grund (2013: 428), the editing of alchemical manuscript texts can be called 'the final frontier' in Middle English (ME) textual editing. There are currently no digital editions of ME alchemical texts, although one is under preparation (Grund 2006). Indeed, there is to my knowledge only one digital edition of alchemical texts from any period: *The Chymistry of Isaac Newton*, presenting Newton's alchemical manuscript material (Newman 2005). There are also very few print editions compared to the vast amount of manuscript material: currently, only nine scholarly print editions of ME alchemical manuscript texts exist (Grund 2013: 431-32, fn. 14-15). The lack of editions may be partly due to alchemical texts having been considered too 'obscure' and 'difficult' to merit editing; alchemical language has a reputation for being vague and laden with metaphors. In general, English-language early scientific texts have not been edited much until fairly recently (*cf.* Pahta and Taavitsainen 2004: 3-4), and disciplines such as alchemy, considered pseudo-scientific in the present day, have been especially neglected.

¹ skmnor@utu.fi.

However, alchemical texts present many intriguing research possibilities. In order for this branch of ME scientific writing to be used by *e.g.* historical linguists, more alchemical texts need to be edited – preferably in a digital form compatible with corpus search tools. This paper will discuss the challenges presented by ME alchemical texts and the ways in which a digital edition can address those challenges. A group of previously unedited and unresearched alchemical manuscript texts will act as a case study, with a focus on the issue of textual fluidity.

Early English alchemical manuscript texts attributed to Roger Bacon

The earliest Latin translations of Arabic alchemical texts appeared in the 12th century (Principe 2013: 51). However, it was only in the 15th century that texts began to be written in English. At first, these were mostly translations or transformations of Latin originals. The 16th and 17th centuries saw a flourishing of alchemical practice in England, and thus a proliferation of new alchemical texts. However, many older texts also were copied; medieval texts may appear for the first time as an early modern manuscript copy.

Pseudepigraphical texts – that is, texts falsely attributed to a famous author – are common in alchemical writing, as in many other types of medieval writing. Roger Bacon (c. 1214-92?) was a scholar interested in many branches of science. He valued the science of alchemy, and produced some genuine writings on the subject (Molland 2004). However, they are outnumbered by the pseudepigrapha: Bacon rose to great fame in alchemical circles, and thus numerous writings were attributed falsely or spuriously to him even centuries after his death.

Among these Pseudo-Baconian texts are several English-language texts. My doctoral dissertation will include a digital scholarly edition of some of them: the focus in my dissertation is on *The Mirror of Alchemy*, extant in seven manuscript copies. The edition will be oriented linguistically, but with the possibility for a general reader/usership as well due to the flexibility of the digital format. Due to the linguistic focus, the edition will be documentary (it also will be accompanied by a reader-friendly best-text edition of one of the copies). Documentary editions are especially important for historical linguists, since the focus on accurate representation of the original and the lack of unsignalled editorial interventions make the edition a better witness of the language of a past age (Lass 2004). My edition will adapt the framework for digital editing of ME texts proposed by Ville Marttila (2014); the TEI-XML guidelines form the basis for the framework. In the course of my doctoral degree, I will mainly aim for a data archive, intended to include the raw XML data and metadata, but not aiming for visual representation of any great degree. Any website presentation will appear only later. The edition is currently in early stages.

However, in this paper I focus on a broader group, featuring many of the English-language alchemical texts attributed to Roger Bacon. Some manuscript versions of *The Mirror of Alchemy* are included in this broader group. Overall, these texts seem to bear little relation to Bacon's genuine writings. The group of Pseudo-Baconian texts focused on here consists of twelve texts from manuscripts dating

from the 15th to the 17th centuries. The manuscripts are located in libraries in Cambridge (Cambridge University Library, Trinity College Library), London (British Library) and Oxford (Bodleian Library). I have transcribed all of the texts; they add up to c. 31,400 words.

Ten of these texts have been divided preliminarily into four groups (A, B, C and D) based on their identification by George Keiser as versions of the same 'work' (Keiser 1998: 3636). The A and D groups consist of only one manuscript text each; the B group consists of five texts, and the C group of three. The material also includes two texts I have found which do not fit Keiser's classification. All the texts are predominantly in English, with occasional code-switching into Latin. The B group has a clear Latin antecedent: *Speculum Alchemiae*, also falsely attributed to Roger Bacon. The B group thus can be identified as versions of *The Mirror of Alchemy*.

Most of the texts are treatises on alchemy, dealing with matters such as the properties of metals and how to prepare the Philosophers' Stone. However, there are also several alchemical recipes.

The issue of textual fluidity and other challenges

Early alchemical texts present several challenges to the editor. Some of these are common to the editing of any manuscript text. For instance, physical challenges such as smudges, stains, faded ink and so on are not a unique problem for alchemical texts. However, some of these physical challenges may be exacerbated by the fact that many alchemical manuscripts actually have been used as workbooks before being placed in archives, and so may contain stains from various substances used in alchemical work (e.g. a small orange-hued stain penetrating several folios in Oxford, Bodleian Library MS Ashmole 1486, Part IV, at least ff. 17r-18v).

In my introduction I mentioned the ambiguity of alchemical language. Metaphors are indeed common, and although there has been plenty of research into alchemical metaphors and imagery (e.g. Abraham 1998), individual texts may well be obscure. Some of the Pseudo-Baconian texts rely on metaphors such as the 'father and mother of all metals' (referring to sulphur and mercury, often conceived of as male and female, respectively; cf. Principe 2013: 78), and other more obscure ones such as 'byardyd blood' (London, British Library, MS Sloane 1091, f. 102r; it is unclear precisely what substance this refers to). There are additional challenges to understanding the texts to be edited: alchemical terminology is not always recorded in the *Oxford English Dictionary* or the *Middle English Dictionary* (cf. Grund 2014), so it can be challenging to uncover the meaning of technical terms (e.g. '*contricyons*', MS Sloane 1091, f. 102v).

There are also challenges relating to representing the original manuscript in the edition; many of these, of course, are common to any editor, but alchemical manuscripts often add to these problems. For instance, alchemical manuscripts often contain extensive marginalia and additions that may be of great interest linguistically and from a historical point of view (e.g. a note in mirrored handwriting in Cambridge, Trinity College Library MS R. 14. 44, in Part IV of the MS, f. 12v). Thus, these marginalia also should be encoded as intrinsic to the

text itself, but presented as separate depending on whether they were added by the original scribe or a later hand. Since there are so few other editions, the possibility for comparison is not yet as fruitful as it could be – however, the Isaac Newton project is an extremely helpful resource for Early Modern English alchemical texts.

Medieval alchemical texts are complex when it comes to textual transmission, as is the case for many other early scientific texts (cf. Grund 2013: 435; Pahta and Taavitsainen 2004: 12–13). Scribes combined sections from various sources and sometimes added their own contributions. Perhaps the chief editorial challenge, thus, is that due to the fluid nature of alchemical texts, it is often difficult to actually define what a certain ‘text’ actually is (cf. Varila 2016; the term ‘work’ is useful in some cases, such as for *The Mirror of Alchemy*, but not in all). This is a challenge if one is considering a printed edition where (for either financial or practical reasons) it is often not feasible to attempt a documentary record of all the possible textual variations.

An example of this textual fluidity can be found in the Pseudo-Baconian D text (MS R. 14. 44, Pt IV, ff. 8v – 14v). According to Keiser (1998: 3636), this text forms a separate group on its own, and thus is not connected to the other texts in his four-part classification. However, my transcription of the D text reveals some definite connections: D contains passages similar to the C group. It also has similar collocations compared to the B group. Keiser’s (1998) work is a manual of scientific writings in ME, and he had to go through a great number of manuscripts in the process. Thus, it is not surprising that extensive research on all the manuscripts was not feasible. In any case, it seems clear that Keiser’s group boundaries are more fluid than they might appear, and should be reconsidered.

The textual fluidity in the D group is evident when comparing it to the C group. One of the texts in the C group (Cambridge, Trinity College MS R. 14. 45, ff. 2r – 4r) contains the following on f. 2r:²

Ask þ^e comyn verkerys þat holde ham soo wyse what ys þ^e
erthe & what ys the vete. þat schall be souyn in the erthe.
*‘ask the common workers that consider themselves so wise: what is the
earth and what is the wheat that shall be sown in the earth’*

The same passage also appears in the other texts in the C group. When compared with a passage from the D text (f. 12r), it can be seen that the two examples are almost identical according to medieval standards:

aske 3e of þese. philisophires. þat holden hem so wyse.
what muste be þ^e whete þat is sowyn in þ^e erthe
*‘ask ye of these philosophers that consider themselves so wise:
what must be the wheat that is sown in the earth’*

2 The transcriptions are by myself. My transcription principles, in brief, are the following: original spelling, line breaks, punctuation, and superscripts have been retained. Abbreviations are expanded in italics. Thorn and yogh are retained.

The only major difference in word choice is ‘comyn verkerys’ ‘common workers’ in the C text, which is ‘philisophires’ ‘philosophers’ in the D text. However, the general similarities are immediately evident. In addition, the passages continue in a near-identical fashion in both texts until D’s explicit, over the space of several folios. There are also other similarities; the texts in MSS R. 14. 44 and R. 14. 45 (as well as the other C texts) have numerous complexities in their textual relationships that cannot be dealt with here. Thus, the latter part of the D text seems to have a similar textual history as parts of the C texts. Should the D text then be included among the C group? Or is it sufficiently different to merit a grouping of its own? Editorial decisions such as these are made very challenging by the textual situation in the manuscripts. However, digital documentary editing can provide many solutions for these issues.

Digital editions for alchemical texts

Digital editing is a useful choice for alchemical texts for many reasons. Considering the issues of textual fluidity, a digital edition is a good solution: because of the lack of issues such as printing costs, a digital edition of alchemical texts can provide all the versions of a text and represent their interrelations in a flexible manner. Digital editions can overcome the issue of ‘too much’ text to be edited: lack of space is not an issue in the digital realm. Thus, multiple versions of an alchemical text/work can be edited and displayed in various ways. In addition, digital editions in website form can make it easier to present complex interrelations of texts. On a website, it would be possible to display links in textual organisation – thus enabling comparison of different versions. Such an edition could also highlight such things as the similarity of passages (*cf.* the example of the C and D groups above).

In addition, digital editions can easily provide multiple representations of alchemical texts with varying degrees of normalisation, thus catering to audiences both scholarly and popular. This makes the texts both 1) accurate representations of historical evidence – full texts, with all the idiosyncrasies of spelling and word choice used by the original scribe – and also 2) accessible for a more general audience, with the possibility of a representation of the edition with *e.g.* normalised spelling. Varying degrees of normalisation – with the possibility of going back to the most accurate representation of the manuscript – are one of the great strengths of digital editions. In one representation, *e.g.* modern punctuation can be added to ease comprehension of the often convoluted syntax of ME alchemical language. However, this should be encoded clearly in the XML as an editorial addition, and should not be part of the default view of the texts.

In conclusion: we need more editions of alchemical texts. The examples presented here are but a drop in the ocean of unedited material. Digital editing is very well suited to alchemical texts. However, a common framework for the documentary editing of alchemical texts is needed. In my doctoral dissertation and the accompanying edition, I hope to present a suggestion for such a framework.

References

- Abraham, Lyndy. 1998. *A Dictionary of Alchemical Imagery*. Cambridge: Cambridge University Press.
- Grund, Peter J. 2006. 'Manuscripts as Sources for Linguistic Research: A Methodological Case Study Based on the Mirror of Lights'. *Journal of English Linguistics* 34(2): 105-25.
- . 2013. 'Editing alchemical texts in Middle English: The final frontier?'. In *Probable Truth: Editing Medieval Texts from Britain in the Twenty-First Century*, edited by Vincent Gillespie and Anne Hudson. Turnhout, Belgium: Brepols, 427-42.
- . 2014. 'The 'forgotten' language of Middle English alchemy: Exploring alchemical lexis in the *MED* and *OED*'. *The Review of English Studies* 65: 575-95.
- Keiser, George R. 1998. *Manual of the Writings in Middle English 1050-1500, Volume 10: Works of Science and Information*. New Haven, CT: Connecticut Academy of Arts and Sciences.
- Lass, Roger. 2004. 'Ut custodiant litteras: Editions, corpora and witnesshood'. In *Methods and Data in English Historical Dialectology*, edited by Marina Dossena and Roger Lass. Bern: Peter Lang, 21-48.
- Marttila, Ville. 2014. 'Creating Digital Editions for Corpus Linguistics: The case of *Potage Dyvers*, a family of six Middle English recipe collections'. PhD dissertation. University of Helsinki, Department of Modern Languages. <http://urn.fi/URN:ISBN:978-951-51-0060-3>. (accessed 27 February 2017)
- Molland, George. 2004. 'Bacon, Roger (c. 1214-1292?)'. *Oxford Dictionary of National Biography*. Oxford University Press. <http://www.oxforddnb.com/view/article/1008>. (accessed 27 February 2017)
- Newman, William R. (ed.) 2005. *The Chymistry of Isaac Newton*. Indiana University. <http://webapp1.dlib.indiana.edu/newton/>. (accessed 27 February 2017)
- Pahta, Päivi and Irma Taavitsainen. 2004. 'Vernacularisation of scientific and medical writing in its sociohistorical context'. In *Medical and Scientific Writing in Late Medieval English*, edited by Irma Taavitsainen and Päivi Pahta, 1-22. Cambridge: Cambridge University Press.
- Principe, Lawrence M. 2013. *The Secrets of Alchemy*. Chicago: The University of Chicago Press.
- Varila, Mari-Liisa. 2016. *In search of textual boundaries: A case study on the transmission of scientific writing in 16th-century England*. Anglicana Turkuensia 31. Turku: University of Turku.

Towards a digital edition of the Minor Greek Geographers

Chiara Palladino¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The purpose of the presentation is to introduce in-progress work for the creation of a digital edition of the corpus of the so-called Minor Greek Geographers. It will focus on issues and opportunities in the formalization and use of geographical data expressed in historical languages such as Greek and Latin.

The definition of 'Minor geographers' has become traditional in scholarship, due to the establishment of a distinction between 'major' geographical encyclopedias of antiquity (Strabo's or Pausanias' works, for example), and 'minor' compilations or opuscles belonging to essentially anonymous authors, of relatively short extent or in fragmentary form, whose coverage was more technical or strictly related to travel.

This ensemble is traditionally indicated as a 'corpus', as the majority of these works were transmitted in two reciprocally connected manuscripts (Pal. Gr. 398 and Par. Suppl. Gr. 443, 9th and 13rd century respectively). These two related *codices* were transmitted together throughout the tradition, and were edited for the first time in 1855-1861 by Karl Müller in the two volumes of the *Geographi Graeci Minores* – that also included works in Latin, derived from Greek originals (Müller 1855; 1861).

However, the texts falling in this definition are varied in terms of sources, chronological, spatial and linguistic coverage, and also in terms of purpose. The corpus includes summaries of astronomy, didactic *compendia* of geography, travelogues, or even poetical *extravaganzas* (e.g. portolans transposed in verses, such as Dionysius' *Periegesis of the inhabited world*). Therefore, they are strategical repositories of information: in a relatively limited textual space, they provide immense variety and density of geospatial data about the ancient world (average

¹ chiara.palladino1@gmail.com.

4-5 place names per line are mentioned in prose works), frequently preserving extended quotations from previous not surviving sources, such as Dicæarchus, Eratosthenes and Hipparchus; their chronological span also offers a wide range of linguistic information on the expression of space.

Despite being certainly outdated, Müller's work offers a panoramic view on a corpus of spatial documents which is currently unique in its completeness and variety. His gigantic critical effort contributed to the creation of a conceptual 'collection' of interrelated works which not only are connected from the codicological point of view, but also represent a geographical 'tradition' of technical and didactic documents which contributed to shape the spatial knowledge and experience of the Graeco-Roman world in the form of a conceptual transmission, a tradition of information from the Classical period to the Byzantine Age, with all the imaginable variations, enhancements, discussions and distortions (Van Paassen 1957). Müller's edition represents, therefore, the starting point of any critical work on ancient geography.

The two volumes have been scanned and are currently available as copyright-free in .pdf format on archive.org and Google Books. In order to extract a machine-readable text from the scans, we used the Lace OCR engine developed by Federico Boschetti and Bruce Robertson (Robertson 2012; Boschetti and Robertson 2013), which is based on Rigaudon, originally developed by Robertson for Polytonic Greek (Robertson 2014), and trained on different fonts used in modern critical editions. While the first tests revealed very positive results for the first volume of the GGM, the scans of the second volume were of very bad typographic quality, and the workflow had to be repeated several times with new scans taken from the physical book. Both the Greek and the Latin translation were OCR'd from the first volume with good accuracy, and have been manually corrected and enriched with basic XML structural tagging. The resulting files can then be used to create a more refined encoding through manual tagging and the creation of a machine-readable citation scheme.

Following the editorial tradition of his time, Müller also provided Latin translations for every text, often copying them from previous single or collective editions, sometimes providing them himself. The availability of a corresponding Latin translation to the Greek text offers an important opportunity to create translation alignment. The alignment editor is in this case provided by Alpheios (<http://alpheios.net/>) through the Perseids platform (<http://www.perseids.org/>), which provides a collaborative annotation environment. The annotations can be exported in XML format, also providing a large corpus of training data for the improvement of automatic cross-language alignment engines.

These steps belong to what should be called a preliminary editorial workflow, aimed at providing machine-readable texts with basic structural features. The inspection of conceptual/semantic phenomena, on the other hand, has to be approached by means of external annotation. The corpus of the Minor Geographers offers a panoramic view on a very specific type of data, namely, geospatial information.

Spatial data in textual sources are essentially based on the notion of ‘place’, which should not be limited to place-names, but provide a fundamental point of start. The basic workflow for machine-readable sources including information about places requires the use of standard references to authority lists, namely, gazetteers. Gazetteers currently represent an authoritative reference to geographical entities, which are stored as Uniform Resource Identifiers (URIs), referred to real geographical coordinates when possible, and also classified according to modern categorization systems. This means that each place is identified uniquely through a stable machine-readable reference, included in a more general authority reference where it can be associated to additional information, such as alternative names, chronological data, disambiguation references, related projects, etc.

In the case of historical languages such as ancient Greek, automatic methods on Named Entities are still very imperfect and computer-based recognition workflows still need to be reinforced with manual intervention, especially in the task of disambiguation and georeferencing. The most important explorative environment where this currently is attempted is provided by Recogito (<http://pelagios.org/recogito/>), the annotation platform created within the Pelagios Project (Barker *et al.* 2015). Recogito was developed in order to allow manual and semi-automatic annotation on geospatial documents in historical languages, to index and georeference place names by referring them to gazetteers, and to allow simple visual representation and data download. It is currently under development in a more advanced version which eventually will allow annotation of an extended variety of named entities (<http://recogito.pelagios.org/>).

Georeferencing and visually representing the data provided by dense geospatial texts gives some additional possibilities in terms of comparative study and global representation of the general spatial concepts of the corpus. Moreover, the principle of collaborative annotation allows the extension of the work to scholarly and students’ projects involving Greek or Roman geographers, for example in the context of seminars dealing with Named Entities data in ancient sources (Bodard and Palladino 2016): some of the authors of Müller’s corpus have already been annotated and georeferenced, and many others are in the process of being annotated by non-experts and scholars (Dionysius’ *Periegesis of the inhabited world*, the *Periplus Ponti Euxini*, *The Greek Cities* by the so-called Heraclides Criticus, Arrian’s *Periplus* and *Indica*, Agathemerus’ *Sketch of Geography* are amongst the most complete in the original Greek version).

Manually verified and georeferenced place-names already provide a database of information which can be used for a variety of purposes: by cross-referencing to Linked Data-based gazetteers such as Pleiades (<http://pleiades.stoa.org/>), for example, additional data are made immediately available, especially in terms of semantic information on place classification. Recogito itself allows for a simple visualization of some of this information, *e.g.* the spatial coordinates as retrieved from georeferencing, boundaries of Roman provinces, and basic information of place frequency indicated by different sizes of pointers. Exported annotations can be used for more refined visualization experiments, by using some of the other

categories of information provided: applications like QuantumGIS (<http://www.qgis.org/>), for instance, allow the elaboration of occurrence-based relations or connections (although the semantic meaning of such relations is not explicit), density and frequency data in the mention of specific categories of places, such as ethnics, water bodies or man-made structures.

The expression of space through specific morphosyntactic constructs is also important for the study of the corpus, as the linguistic encoding of directions and orientation has a relevant cognitive importance (Levinson 2003; Thiering 2014): this is approached through treebanking, linguistic annotation of morphology and syntax (Celano n. d.).

Geospatial documents generated by premodern societies challenge our spatial perspective in a variety of other ways, and they still provide much meaningful information that should be treated with a specific formalization strategy in order to be fully inspected and understood. When dealing with ancient geography, it has to be acknowledged that premodern societies were not map-based, as they considered cartographic representation of space as mathematical-philosophical abstraction, while concrete traveling and spatial practice were related to descriptive geography in word-form (Janni 1984; Brodersen 2003). Spatial practice was shaped through a peculiar system of knowledge, which made limited use of graphical representation, but was built as a chain of concepts and specific components, functional to empirical experience (Geus and Thiering 2014). These components can be summarized as relational descriptions of various types: the main ones being distances, definition of topological/conceptual categories, and systems of orientation, either formalized or non-formalized.

The corpus of the Minor Geographers offers a variety and density of geospatial information which provides a suitable quantity of documents representing the way Graeco-Roman societies described and experienced the world through specific concepts, in a radically different way from our perceived cultural paradigm. On the other hand, it challenges the limitations of current GIS technologies and editorial workflows, which are not apt to represent non-standardized ways of connection between spatial entities, with all their semantic peculiarities, in a consistent way. For example, it is currently impossible to refer to any authority concerning types of orientation systems, or ancient classifications of places (which tend to be different from the modern ones), and especially relations between spatial entities in general. What is still missing from the picture is, therefore, an authoritative vocabulary retrieved from textual data as they appear in ancient sources, not only classifying places, but especially providing additional relational and semantic information.

References

- Barker, Elton, Rainer Simon, Leif Isaksen and Pau De Soto Cañamares 2015 'Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito.' *e-Perimtron* 10. 2: 49-59.
- Bodard, Gabriel, and Chiara Palladino. 2016. 'Named Entity Recognition: SNAP and Recogito', *SunoikisisDC-2016* online seminar. <https://github.com/SunoikisisDC/SunoikisisDC-2016/wiki/Named-Entity-Recognition:-SNAP-and-Recogito-%28February-24%29>.
- Boschetti, Federico, and Bruce Robertson. 2013. 'Lace: Greek OCR.' <http://hempl.mta.ca/lace/>.
- Brodersen, Kai. 2003. *Terra cognita: Studien zur Römischen Raumerfassung*. Olms: Hildesheim, repr.
- Celano, Giuseppe G. A. n. d.'Guidelines for the Ancient Greek Dependency Treebank 2. 0', https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md.
- Geus, Klaus, and Martin Thiering (eds). 2014. *Features of Common Sense Geography: implicit knowledge structures in ancient geographical texts*, Berlin-Münster-Wien-Zürich-London: LIT Verlag.
- Janni, Pietro. 1984. *La mappa e il periplo: cartografia antica e spazio odologico*. Roma: Brentschneider.
- Levinson, Stephen C. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: University Press.
- Müller, Karl. 1855. *Geographi Graeci Minores, Volumen Primum*. Paris: Didot.
- . 1861. *Geographi Graeci Minores, Volumen Secundum*. Paris: Didot.
- Robertson, Bruce. 2012. 'Optical Character Recognition of 19th Century Polytonic Greek Texts – Results of a preliminary survey.' <http://www.hempl.org/RobertsonGreekOCR/>.
- Robertson, Bruce. 2014. 'Rigaudon.' <https://github.com/brobertson/rigaudon>.
- Thiering, Martin. 2014. *Spatial semiotics and spatial mental models: figure-ground asymmetries in language*. Berlin: De Gruyter.
- Van Paassen, Christian. 1957. *The Classical tradition of Geography*. Groningen: Wolters.

Digital editions and materiality

A media-specific analysis of the first and the last edition of Michael Joyce's *Afternoon*

*Mehdy Sedaghat Payam*¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

In this paper, the early development of hypertext fiction will be approached from the perspective of the materiality of the digital text, with the purpose of demonstrating the extent to which the materiality of the digital medium has affected editing Michael Joyce's *Afternoon*, which as the first work of hypertext fiction is arguably the most discussed work of the early 1990s. Writing a novel in the new medium and presenting it as a work of fiction surely must have required a significant amount of (pre)meditation about narrative, the new medium and the way its materiality should or should not play a significant part in the narrative, and this is what Joyce confesses to have done during the years preceding the writing of *Afternoon*. The main reason which this novel has been chosen for this research is that *Afternoon* in various ways reveals how it changes when its material support as well as its reading and writing software, Storyspace, go through various updates. Those updates practically make a floppy disc designed to be read through Storyspace 1.0 on a Mac LC, unreadable on an iMac with Storyspace 1.3.0. The stand-alone feature of the works of hypertext fiction means that their material support needs to be updated anytime that a new technology, or the upgrades to the previous ones, change the electronic media ecology. This has already happened once when Storyspace 1 was upgraded to Storyspace 2, which according to its developer made it a completely new computing environment. The current Storyspace available on the website of Eastgate System (late 2015) is Storyspace 2.5 for Mac OS X, and Storyspace 2.0 for Windows. During the last twenty years the medium of the computer and its operating systems were developed further and further by their manufacturing

¹ ms79payam@yahoo.com.

companies. This made Joyce (and his publisher) develop *Afternoon* and edited it for the new platforms. For the contemporary reader in late 2015, apart from the web ported version only the sixth edition (for both Mac and Windows users) is commercially available.

As both Kirschenbaum (2008) and Harpold (2009) have demonstrated, there are several *Afternoons*. Just as during the last twenty years the medium of the computer and its operating systems were developed further and further by their manufacturing companies, so did Joyce (and his publisher) develop *Afternoon*. Harpold has identified sixteen versions for *Afternoon* up to 2007, six for Macintosh computers and six for Windows, two translations (into German and Italian), a web porting, and a print version. The last two are only a selection of the lexias (fifteen lexias for the web version, and ten for the print copy of *Postmodern American Fiction: A Norton Anthology*). For the contemporary reader in late 2015, apart from the web ported version only the sixth edition (for both Mac and Windows users) is commercially available, and the previous editions are already collector's items.

Although *Afternoon* as the first hypertext novel has primary importance for this paper, its history, development and the myriad of changes that it has undergone in each edition will not be studied here.² Instead, this paper analyzes the first commercial edition of the novel and the sixth edition to show how and in what ways the materiality of this novel has been modified for different platforms and operating systems. Moreover, it will demonstrate how all these features are dependent on and a function of the interplay between this novel's physical characteristics and its signifying strategies.

The materiality of both editions of *Afternoon* should be studied in at least three ways, two of which are forensic and one of which is formal materiality.³ For the first commercial edition, the first kind of forensic materiality refers to the data storage medium (3.5 inch floppy disks) on which it was sold, and the second kind of forensic materiality refers to the machine which retrieves information from the storage medium and makes reading *Afternoon* possible (a Mac LC here). The formal materiality of *Afternoon* refers to the software (Storyspace) by which this novel has been written and presented to the audience. All these three at the same time provide various opportunities and create several limitations for the early works of hypertext fiction writers. In addition to all these, packaging and the way the floppy diskettes have been sold by their publisher (Eastgate Systems) creates another level of materiality which will be discussed below. Michael Joyce who was one of the developers of Storyspace states that it took him four years to figure out how to make it a useful tool in writing novels.

2 Terry Harpold (2009) and Mathew Kirschenbaum (2008) have written insightful chapters on these issues in their books. Reading their books can be very informative for readers who are interested to explore these fields.

3 Formal materiality affects the form of the files and not the substance of the computer. In other words, formal materiality is an emphasis on the manipulation of symbols not the matter. It is experienced as buttons on the screen or a blank page that is filled by users while writing on a MS Word Document. Forensic materiality on the other hand, refers to all the physical elements (both microscopic and macroscopic) used in the computer (C.f. Kirschenbaum 2008).

Yet a fairly common reaction to hypertext and hypermedia systems in product reviews, in technical literature, and among everyday users of these tools is the one expressed by Jeffrey Conklin in his still definitive 1987 article, 'One must work in current hypertext environments for a while for the collection of features to coalesce in a useful tool.' This is a kind way to say that you have to figure out what to do with these things. I have spent much of the last four years figuring out exactly that. As a code-developer of Storyspace, I have approached what to do with the things as a design question; as a fiction writer seeking to work in a new medium, I have approached it as an artistic question, and, as a teacher, a practical, pedagogic, and sometimes a political one.

(Of Two Minds 39)

Thus, it can be said that many of the features of *Afternoon* already had been thought about by Joyce, and he created that novel with a comprehensive knowledge of the capabilities of the medium and both layers of its materiality, the computer and its new range of abilities (which required its users to have a certain degree of computer literacy to be able to use it) and Storyspace as the authoring tool for *Afternoon*.

Afternoon as the first work of hypertext fiction, is one of the most discussed works of Storyspace school. From Landow (1994) to Bell (2010) many scholars have discussed various features of this work (closure and possible stories (Douglas, 2001), freedom of the reader (Aarseth 1997), and narrative (Ryan 2006). However, apart from Kirschenbaum (2008), and Harpold (2009), almost none of the books and essays have discussed its materiality, and how changes in different published versions of *Afternoon* modify the experience of reading this work.

The medium used here to analyze the first commercial edition (instead of a simulated program) has been an actual Mac LC with forty MB Hard Drive with 6.0 Mac OS and two MB RAM. The CRT monitor for this Mac was a fourteen inch ViewSonic E641 which has a display resolution of 512x384 pixels at eight-bit color. *Afternoon* was published on a three and a half inch floppy disk which had the capacity limit of 720 Kilobytes and during the early 1990s was the most popular portable storage medium for electronic data. This storage medium imposes its own limitations (in addition to the opportunities it provides) for publishing a work of hypertext fiction. The first (and probably the most important) limitation imposed on the authors by this medium was its small storage capacity which discouraged the authors who wanted to sell their works through Eastgate Systems (and probably designers of the software) from creating large-scale works.⁴ This could have been one of the reasons why *Afternoon*, like many other works which follow it, is mainly script-based and has minimal graphics. The only picture on the opening lexia which shows three characters is small and blurry.

4 The second release of Storyspace, *King of Space* (1991) by Susan Smith was published on three 3.5 floppy disks, and *Uncle Buddy's Phantom Funhouse* (1992) by John McDaid was published on five 3.5 floppy disks (and comes in a box which is totally different from the standard packaging of other Storyspace releases). However, these are the exceptions and most other works released by Eastgate Systems were either published on one or two 3.5 floppy disks.

The version of *Afternoon* which is studied here comes in a jacket which, in an obvious attempt to remediate books, functions as the cover for the novel and, similar to its print equivalents, has the name of the novel and the author on the front. A short introduction about Michael Joyce along with a selection from the reviews of other famous novelists and scholars is printed on its back. Inside the cover, in addition to the floppy diskette of *Afternoon*, there is an actual printed booklet.⁵ This eight-page booklet has eight sections: a) *Afternoon, a story* (with more selections from the reviews), b) Getting Started, c) Reading *Afternoon*, d) The Tool Bar, e) Questions?, f) About the author, g) License Agreement, h) Limited Warranty and Disclaimer. Sections c and d (in slightly altered forms) are repeated in the digital version of the story under the lexias of ‘hypertext’ and ‘read at depth’ and precede the story. Sections a, e, f, g, and h are the standard sections which can be found in almost all contemporary print novels, and adding them to the booklet here is a further attempt at the remediation of the material form of the print novel. It is not clear whether having a printed booklet in packaging has been the decision of Storyspace designers or the publishers (Eastgate Systems), but it clearly aims to further justify Joyce’s claim that *Afternoon* is a novel.

One of the main challenges for Joyce, as the first hypertext fiction writer, was finding an electronic equivalent for the page and how to move to other pages of the novel on the computer screen. In order to address this issue Joyce chose to use (hidden) links and have every lexia fit to the screen, so that the novel could partly have a visual similarity with the book. Another issue which has been taken into account in the physical organization and design of *Afternoon* is the bookmarking capability of the novel, which is shown in the program’s asking to save the reading path before exiting the program while you open it next time for a new reading.⁶ This feature allows the machine to automatically trace the reading sequence of every reading and stresses the procedural aspect of the reading. In a print novel, a reader may use a custom-made bookmark or a piece of paper to bookmark his reading path, and after the novel is finished, the reader does not use the bookmark anymore. However, even if you have visited almost all 539 lexias of the story and have made an intensive study of *Afternoon* like Douglas (2001), the machine still asks you whether you want to ‘save the current place in your reading of *Afternoon*,’ a request which implicitly notes the never-ending nature of reading this novel. No reading is complete and the physical organization and design of this novel stresses this repeatedly. This bookmarking feature might seem trivial at the beginning, but it is one of the features of the software which has been thought seriously about (and later developed) by the developers of Storyspace. Bookmarking in addition to implying kinship with print novels (which cannot be read in one sitting) also implies the existence of the guard-fields. You cannot simply pass through the maze of *Afternoon* in one reading. It is a place that you have to return to (probably every

5 This booklet hints at strength of the print tradition which exists and leaves its mark and trace on even the works which ironically were hailed by some critics as the nemesis of print.

6 The accompanying booklet of *Afternoon* pinpoints this feature: ‘Each time you open *Afternoon*, you can choose whether to begin a new reading or resume previous reading. If you resume a previous reading, you begin at the place where you stopped reading previously.’

afternoon after work) and that is why you need bookmarks to help you explore various corners of it.

Afternoon's 2001 edition for Windows and Mac comes in a CD with a glass cover and the booklet which accompanied the first commercial edition has been downsized to a small photo. Moreover, the information on the booklet has been added as separate file which appears in the same folder where *Afternoon* is. There is also a text file in which the copyright information of *Afternoon* appears. Several other navigation options have been added to this version as well. For instance 'locate writing space' from the Navigate tab allows the reader to search among the lexias. Apart from all this, the 2001 edition has an important statement by Joyce. Under the Storyspace tab in the software, there are a couple of keywords which differ in Windows and Mac versions. The Keywords in Windows are *fragments, moaning, poetry, wall, winter, yesterday*, but the Mac version has some additional keywords which include *black, blue, cyan, green, magenta, and red*. This becomes even more interesting when we search these colours in the 'find text' search option of the novel (under the drop-down Navigate tab), and find out that some of these colors (such as cyan, and magenta) have not been mentioned in the novel at all. Cyan, Magenta, Yellow and Black are the basic colors used for printing color images. On the other hand, Red, Blue and Green are used for creating images on the computer screens. The first obvious interpretation of this feature is that, similar to the creation of colors, different stories are created when different lexias of this novel are placed alongside each other. The second interpretation is that since cyan, magenta, and yellow are 'subtractive' colors, they get darker as you blend them together, and ultimately create the color black. On the other hand, red, green, and blue produce the color white when they are blended together. Since the only colors used in this novel are black on the white background, what Joyce means here is that the whole range of colors are available in his novel, and his work is a blend of all these colors which have been used for creating images on the computer screen and printing them on paper. What this means in textual practice is that his story incorporates the elements of print and computing media and combines them together to achieve a kind of synthesis (or compromise) between these two; the long tradition of the novel and the new medium of the computer.

These changes make a bold statement about the relationship between each edition and its material support which highlights the role of editing for each instantiation of the same work. A work of digital fiction has to evolve alongside the upgrades of the each medium in it is instantiated. These upgrades appear in minor edits on some parts either within the linguistic section of the novel (Harpold 2009; and Kirschenbaum 2008) or in the materiality of the novel, so that even a different platform (Mac OS) offers a different reading of the same work.

References

- Aarseth, Espen J. 1997. *Cybertext: Perspectives on Ergodic Literature*. Baltimore, MD: Johns Hopkins University Press.
- Bell, Alice. 2010. *The Possible Worlds of Hypertext Fiction*. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.
- Douglas, J. Yellowlees. 2000. *The End of Book – or Books Without End? – Reading Interactive Narratives*. Ann Arbor: University of Michigan Press.
- Geyh, Paula, Fred G. Leebron, and Andrew Levy. 1998. *Postmodern American Fiction: A Norton Anthology*. New York: W. W. Norton.
- Harpold, Terry. 2009. *Ex-foliations Reading Machines and the Upgrade Path*. Minneapolis: University of Minnesota Press.
- Joyce, Michael. 1990. *Afternoon; a story*. Floppy Disc. Watertown, MA: Eastgate Systems.
- . 2001a. *Afternoon; a story*. Compact Disc. Watertown, MA: Eastgate Systems.
- . 2001b. *Moral Tales and Meditations: Technological Parables and Refractions*. Albany: State University of New York Press.
- . 1995. *Of Two Minds: Hypertext Pedagogy and Poetics*. Ann Arbor: University of Michigan Press.
- . 2000. *Othermindedness: The Emergence of Network Culture*. Ann Arbor: University of Michigan Press.
- Kirschenbaum, Matthew G. 2008. *Mechanisms: New Media and the Forensic Imagination*. Cambridge MA: MIT Press.
- Landow, George P. 1997. *Hypertext 2.0*. Baltimore: Johns Hopkins University Press.
- Ryan, Marie-Laure. 2006. *Avatars of Story*. Minneapolis: University of Minnesota Press.

Challenges of a digital approach

Considerations for an edition of Pedro Homem de Mello's poetry

Elsa Pereira¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Pedro Homem de Mello (1904-1984) is one of the relevant lyric poets in 20th century Portuguese literature, whose works were subject to extensive authorial revision. This presentation argues that structural variance, as observed in the poet's compositions, can be visualized best through an electronic edition, and that the affordances of the medium contribute to reconceiving the concept of text, as a fluid entity that is subject to expansion, through its divergent manifestations. With that in mind, it will present examples where this kind of variation occurs, and point out a few challenges of going digital.

The first thing we have to observe is that Pedro Homem de Mello started his writing career during the Modernist period. Hannah Sullivan, in her book *The Work of Revision*, already identified obsessive authorial revision with 'the legacy of high modernism' and the multistage process that characterized the 'print culture that nourished it.' She claims that 'Modernism writers (...) used revision' not only through the kind of lexical substitution that aims to refine the expression, 'but to make it new through large-scale transformations' of length and structure (Sullivan 2013, 22). This also could be related to what Tanselle classified as *vertical revision* – that is, a rewriting that metamorphoses a text and seems 'to make a different sort of work out of it' (Tanselle 1990, 53).

There are many examples of vertical revision in Mello's poetry. Sometimes, the relationships between versions are obvious, as they involve mainly excision, extension or substitution of some parts of the poem, while the core of the composition remains the same. Other versions, however, seem to have more of

¹ epereira@net.sapo.pt.

‘a filial relationship rather than one of direct descent’ (Sullivan 2013, 18). This is the case, for example, with two versions entitled ‘Saudade’ (Mello 1961, 49) and ‘Apolo’ (Mello, BNP, E14, cx. 22). Altogether, only four half lines are shared, and those lines do not even appear in the same position. But there is more: these are just two of the nineteen witnesses assembled from a larger cluster of texts that have been identified so far. If we compare them all, we will find several different versions, with extreme variation from one another.

In such cases, we perhaps should ask what degree of variation creates a new version, and if differing versions can be distinguished (and edited) as separate texts. Peter Shillingsburg already has asked these questions and identified four criteria to measure variation: time, content, function, and material (Shillingsburg 2000, 69). The second question remains especially problematic, though. Some critics consider that a work may be regarded as new whenever a certain degree of continuity is not preserved from one stage to the next (McLaverly 1991, 137). Others argue that, no matter how different they may seem to be, versions ‘can never be revised into a different work’ (Bryant 2005, 85), and in no circumstance can be unlinked from one another. John Bryant is included in the second group and recommends that we edit multiple versions at once, mapping out the variation from one stage to another, and enabling users ‘to lead themselves along those paths’ (Bryant 2005, 123). Instead of an appended apparatus ‘that places that data in footnotes or endnotes’ (Bryant 2005, 127), editorial strategies must be devised to combine text and apparatus, thus enticing ‘readers into the pleasures of the fluid text’ (Bryant 2005, 133).

This is somehow difficult to achieve in a printed book. Although the idea of an integrated apparatus has been considered since Zeller’s proposal in the early 1970s (and later put into practice in Gabler’s 1984 edition of the *Ulysses*), it is extremely difficult for readers to visualize variation through coded data, when versions are not formally similar. Alternatives should then be considered. Sometimes a facing-page edition can be an option, when two versions with structural variance are involved. However, the material constraints of the book offer limited possibilities when multiple versions coexist.

As Bryant admits, fluid-text editions ‘can best be realized, perhaps *only* realized, through the extraordinary hypertextual features of the electronic medium’ (Bryant 2005, 145). Yet, an electronic approach could be undertaken in several ways, from static editions to what Peter Robinson envisioned as dynamic repositories of textual knowledge (Robinson 2005).

In the first case, the electronic medium comes up as a mere publication platform, which relies on the hyperlinked architecture to relationally organize reading text, emendations, apparatus, and annotations. Because these sections are assembled independently, users still have to reconstruct the layers of the text themselves, as happens with any print edition. While that is, of course, not a bad thing, it offers limited possibilities to deal with texts where instability and versioning are as strongly governing principles as is the case with these poems.

That brings us to a new editorial paradigm, which is based on the functional integration of text and apparatus, so that new patterns and relationships may be rearranged through automatic processing. The most obvious advantage is the

ability to compare multiple versions, at the user's discretion, and explore diachronic revision, with no single version being privileged hierarchically over another.

In order to make this possible, syntax and markup language become an essential part of the design of the edition. This is where the Text Encoding Initiative (<http://www.tei-c.org>) comes into play, as a comprehensive set of guidelines for the representation of texts in digital form, which aims to empower processability and visualization tools.

While, according to a recent study only 37% of the electronic editions Franzini *et al.* (2016) examined follow TEI (mainly because editors need to make their own customizations, in order to address specific phenomena), the number of projects adopting the guidelines 'is gradually increasing', perhaps as a consequence of its 'systematic growth and improvement' (175). This standardised implementation fosters interoperability, which is important not just from a maintenance perspective. It also allows tools to be shared by several similar projects around the world, facilitating one of the most demanding stages in the creation of a digital edition: the design of the graphic user interface. Since modern authors usually display intricate revisions in their texts, sophisticated software is required to visualize the underlying encodings in a flexible and dynamic environment. Developing such tools from scratch would involve extensive technical support, which in no way is available to individual projects such as this. Fortunately, by adopting the TEI guidelines, we can resort to open-source software with all-round tested solutions.

Among the freely available resources, there is one that generally suits the goals of this project and the specificity of the challenges raised by our author's poetry. It is the Versioning Machine (v-machine.org) – an interface for displaying multiple versions of text encoded with the TEI guidelines, which originally was conceived in 2000 by Susan Schreibman and since has been used in a number of projects with similar cases of vertical revision. Its current version 5.0 was released early in 2016 and is HTML5 compatible. This means that it has been developed to suit texts with multimedia requirements, such as Pedro Homem de Mello's poems adapted to fado, allowing image and audio files to be incorporated and aligned with the text.

Through its hypertextual architecture and a TEI-P5 conformant schema, the Versioning Machine is appropriate for a genetic-critical approach, favouring a text-centric view. Thus, emendations may be added to the transcriptions (using the <choice>, <sic> and <corr> elements), while the representation of the writing process is achieved by a parallel display of successive versions, enriched by in-document sub-stratification. Furthermore, a series of pop-up notes also may be encoded, in order to assemble para-genetic materials that somehow are related to the compositional history of the text.

The Versioning Machine's interface allows users to critically engage with the dynamics of revision, by comparing chunks of text, alongside word-by-word or line-by-line comparisons. These may occur in direct correspondence, or may involve different lines of each version.

This compare and contrast logic (that manifests itself across multiple versions at the same time) is only possible thanks to the underlying encoded apparatus. The VM 5.0 supports two different methods, described in the TEI-P5 critical

apparatus tag set, which can be chosen according to the specificity of the text. While parallel-segmentation is the most straightforward approach, it does not easily deal with overlapping relationships of elements and structures. To get around this challenge, the VM 5.0 has built in the ability to encode documents with the location-referenced encoding method, which associates several lines with corresponding text by using the same value for the <loc> attribute within the <app>s that need to be assembled.

The visual rendering is that, by mousing over any of the lines within the same location value, all <rdg>s inside correlating <app>s will be highlighted in the HTML interface, providing a dynamic representation of compositional or revisional stages, which is substantially different from the print counterparts.

Where conventional book editions privilege a stable reading text, fully digital editions like this favour a continual textual flow that is mutable over time, since multiple assembled versions dialogue through the text's subtle fluidities. Such a heuristic display of versions allows for non-linear reading paths, and is particularly significant when dealing with authors from the Modernist period, because, as we have seen with Hannah Sullivan, they revised 'more frequently, at more points in the lifespan of the text, more structurally and experimentally' (Sullivan 2013, 22). Therefore, the decision to edit Pedro Homem de Mello's poems electronically is not just a matter of representation. It actually 'conveys and embodies a' different, 'pluralistic notion of text' (Sahle 2016, 30-31).

References

- Bryant, John. 2005. *The Fluid Text: a Theory of Revision and Editing for Book and Screen*. 4th ed. Ann Arbor: The University of Michigan Press.
- Franzini, Greta, Melissa Terras, and Simon Mahony. 2016. 'A catalogue of digital editions.' In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew J. Driscoll, and Elena Pierazzo, 161-182. Cambridge: Open Book Publishers.
- McLavery, James. 1991. 'Issues of identity and utterance: an intentionalist response to 'textual instability''. In *Devils and Angels: Textual Editing and Literary Theory*, edited by Philip Cohen, 134-151. Charlottesville-London: University Press of Virginia.
- Mello, Pedro Homem de. n. d. 'Apolo.' Typescript. BNP (Biblioteca Nacional de Portugal), E14, cx. 22.
- . 1961. *Expulsos do Governo da Cidade*. Porto: Livraria Galaica.
- Robinson, Peter. 2005. 'Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future?' *Digital Medievalist* 1.
- Sahle, Patrick. 2016. 'What is a Scholarly Digital Edition?' In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew J. Driscoll, and Elena Pierazzo, 19-39. Cambridge: Open Book Publishers.
- Shillingsburg, Peter L. 2000. *Resisting Texts: Authority and Submission in Constructions of Meaning*. 4th ed. Ann Arbor: The University of Michigan Press.
- Sullivan, Hannah. 2013. *The Work of Revision*. Cambridge-Massachusetts: Harvard University Press.
- Tanselle, G. Thomas. 1990. *Textual Criticism and Editing*. Charlottesville: University Press of Virginia.

The born digital record of the writing process

A hands-on workshop on digital forensics, concepts of the forensic record and challenges of its representation in the DSE

Thorsten Ries¹

Workshop tutored at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

In its first part, the workshop gave a hands-on introduction to digital forensic analysis of hard drives for born digital traces of the writing process with different constructed case scenarios. The hands-on experience served as a foundation for a moderated group discussion about how the specific materiality of the digital historical record can be read in philological terms of the critique génétique, how this changes our ideas about text production and consequently the requirements and understanding of representation of the genetic digital born record in a documentary or genetic DSE.

The hands-on workshop introduced participating archivists, philologists and researchers from the humanities to forensic imaging of hard drives, inspection and analysis of forensic images with the forensic Linux distribution Caine Linux 7.0. Two phases of analysis of the process have been covered during the workshop: a) forensic imaging, triage and preservation of hard drives in the archive and b) philological recovery of textual versions of a writing process from a digital forensic image (mounting, inspection of temporary files, undelete, file carving, drive slack analysis, timeline analysis, grep, use of bulk_extractor) and by low-level inspection of files (fast save artefacts, RSID-tags). While doing so, the participants have been introduced to typical use case effects such as file fragmentation in the evidence (binary formats, zip container). Other scenarios, e.g. cloud services, have been

¹ thorsten.ries@ugent.be.

addressed briefly. To avoid legal issues, participants worked with forensic images created for training purposes with Christian Moch's *Forensig2* forensic image generator (Moch 2009).

The guided exercises performed during the workshop led to lively discussions about the preservation of born digital heritage, digital forensic analysis and the representation of digital forensic findings in genetic scholarly editions.

References and Background Reading

- Cohen, Fred. 2011. 'Putting the Science in Digital Forensics.' *Journal of Digital Forensics, Security and Law*, Vol. 6. 1: 7-14.
- Chun, Wendy Hui Kyong. 2008. 'The Enduring Ephemeral, or the Future Is a Memory.' *Critical Inquiry* 35.3: 148-171.
- Duranti, Luciana. 2009. 'From Digital Diplomats to Digital Records Forensics.' *Archivaria* 68: 39-66.
- Duranti, Luciana; Barbara Endicott-Popovsky. 2010. 'Digital Records Forensics. A New Science and Academic Program for Forensic Readiness.' *ADFSL Conference on Digital Forensics, Security and Law* 2010: 109-122. Accessed August 22 2015. <http://arqtleufes.pbworks.com/w/file/fetch/94919918/Duranti.pdf>.
- Emerson, Lori. 2014. *Reading Writing Interfaces. From the Digital to the Bookbound*. University of Minnesota press.
- Kramski, Heinz Werner and Jürgen Enge. 2015. 'Arme Nachlassverwalter...' Herausforderungen, Erkenntnisse und Lösungsansätze bei der Aufbereitung komplexer digitaler Datensammlungen.' In *Von der Übernahme zur Benutzung. Aktuelle Entwicklungen in der digitalen Archivierung. 18. Tagung des Arbeitskreises 'Archivierung von Unterlagen aus digitalen Systemen' am 11. und 12. März in Weimar*, edited by Jörg Filthaut: 53-62. Weimar: Thüringisches Hauptstaatsarchiv.
- Garfinkel Simson, Paul Farrell, Vassil Roussev and George Dinolt. 2009. 'Bringing Science to Digital Forensics with Standardized Forensic Corpora.' *DFRWS* 2009: 2-11.
- Garfinkel, Simson. 2010. 'Digital Forensics Research: The Next 10 Years.' *DFRWS* 2010: 64-73.
- Gitelman, Lisa. 2014. *Paper Knowledge: Toward a Media History of Documents*. Duke University Press.
- Goddard, Michael. 2014. 'Opening Up the Black Boxes. Media Archaeology, 'Anarchaeology' and Media Materiality.' *New Media and Society* 17. 11: 1761-1776. DOI: 10.1177/1461444814532193.
- Hiller, Moritz. 2013. 'Diskurs/Signal (I). Literaturarchive nach Friedrich Kittler.' In *Mediengeschichte nach Friedrich Kittler* (Archiv für Mediengeschichte 13): 147-156. Paderborn, München.
- Hiller, Moritz. 2014. 'Diskurs/Signal (II). Prolegomena zu einer Philologie digitaler Quelltexte.' *Editio* 28: 193-212.
- John, Jeremy Leighton. 2012. 'Digital Forensics and Preservation.' *DPC Technology Watch Report* 12-03 November 2012. Digital Preservation Coalition 2012. Accessed 22 August 2015. DOI: 10.7207/twr12-03.

- Kirschenbaum Matthew. 2006. *Mechanisms. New Media and the Forensic Imagination*. Cambridge, London: MIT Press.
- Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine (eds). 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources Washington.
- Kirschenbaum, Matthew G. 2013. 'The. txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary.' *Digital Humanities Quarterly* 7.1. Accessed 28 February 2015. <http://www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html>.
- Kirschenbaum, Matthew, and Doug Reside. 2013. 'Tracking the changes. Textual scholarship and the challenge of the born digital.' In *The Cambridge Companion to Textual Scholarship*, edited by Neil Freistat and Julia Flanders: 257-273. Cambridge: Cambridge University.
- Kirschenbaum, Matthew. 2014. 'Operating Systems of the Mind. References After Word Processing.' *The Papers of the Bibliographical Society of America* 101. 4: 381-412.
- Kirschenbaum, Matthew. 2016: *Track Changes. A Literary History of Word Processing*. Cambridge: Harvard University Press 2016.
- Kramski, Heinz Werner. 2013. 'Ordnungsstrukturen von der Floppy zur Festplatte. Zur Vereinnahmung komplexer digitaler Datensammlungen im Archivkontext.' Beiträge des Workshops 'Langzeitarchivierung' auf der Informatik 2013 am 20. 09. 2013 in Koblenz. Frankfurt am MaIn nestor Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit *Digitaler Ressourcen für Deutschland* 2014 (= nestor edition -Sonderheft 1): 3-13. <http://files.d-nb.de/nestor/edition/Sonderheft1-Informatik2013.pdf>.
- Moch, Christian. 2009. *Automatisierte Erstellung von Übungsaufgaben in der digitalen Forensik*. PhD Diss., Erlangen: Nürnberg.
- Reside, Doug. 2011. 'No Day But Today:' A look at Jonathan Larson's Word Files. *NYPL Blog*, April 22. <http://www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files>.
- Reside, Doug. 2011. 'Last Modified January 1996.' *The Digital History of RENT. Theatre Survey* 52.2: 335-340.
- Reside, Doug. 2012. 'Digital Genetic Criticism of RENT.' *Talk Digital Humanities 2012 in Hamburg*. Accessed 22 August 2015. <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/digital-genetic-criticism-of-rent.1.html>.
- Reside, Doug. 2014. 'File Not Found: Rarity in an Age of Digital Plenty.' *RBM: A Journal of Books, Manuscripts, and Cultural Heritage* 15.1: 68-74.
- Ries, Thorsten. 2010. 'die geräte klüger als ihre besitzer' Philologische Durchblicke hinter die Schreibszene des Graphical User Interface. Überlegungen zur digitalen Quellenphilologie, mit einer textgenetischen Studie zu Michael Speiers 'ausfahrt st. Nazaire'. *Editio* 24. 1: 149-199.
- Ries, Thorsten. 2016. 'Das digitale dossier génétique. Überlegungen zu Rekonstruktion und Edition digitaler Schreibprozesse anhand von Beispielen aus dem Thomas Kling Archiv.' In *Textgenese und digitales Edieren. Wolfgang*

Koeppens 'Jugend' im Kontext der Editionsphilologie, edited by Katharina Krüger, Elisabetta Mengaldo and Eckhard Schumacher: in press. Berlin *et al.*: de Gruyter.

Vauthier, Bénédicte. 2014. 'La critique génétique à l'épreuve du numérique. El Dorado (2008) de Robert Juan-Cantavella.' *Passim. Bulletin des Schweizerischen Literaturarchivs* 14: 6-7.

Enduring distinctions in textual studies¹

Peter Shillingsburg

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

I was asked to offer a brief retrospective on textual studies. I have been in textual studies since 1966. I first confronted European textual studies seriously in 1999, at a conference of German textual scholars in The Hague. It was an experience I did not understand immediately. Words like authority and mixed authority, scholarly edition, emendation, copy-text, and even apparatus all had different meanings for me because the European and American textual traditions entailed unfamiliar values and nuances.

The history of textual study is punctuated with new insights, but there are enduring theoretical concepts as well. They include basic distinctions. One distinction is that between the material document and the texts inscribed on them. Is a document a unity, where the material and the symbolic are an undivided whole? Or can the inscribed text be replicated endlessly in other documents without violating the unity of being? We have labelled these two ways of looking as the distinction between the bibliographic code (documentary), and the lexical code (symbolic).

Another distinction is between visible printed text as found in a document and the verbal conceptual content or experience of the literary work which the text represents for, or stimulates in, readers. Documents provide textual scholars with sufficient difficulties that we sometimes ignore the conceptual works and mental experiences of the works. I doubt, however, that many of us ever forget that, as Emily Dickinson said, there is no frigate like a book to carry us away from our

1 This is the summary of an address Peter Shillingsburg gave at the joint ESTS/DiXiT convention in Antwerp on the occasion of receiving an award for his work in textual studies by the European Society for Textual Scholarship. A full version is available at <http://sunrisetc.blogspot.com/>. It will not be published elsewhere.

present time and place (Dickinson 1999). Like John Keats, we have boarded those frigates and ‘travell’d in the realms of gold, / And many goodly states and kingdoms seen’ (Keats 1919). With Keats we all have looked into Chapman’s Homer and seen islands. We see through the text to the imagined world. Without that, we probably would not be discussing bibliography, textual criticism, or editorial theory. As textual critics, we look *at* the books, as well as *through* them. It is important not to lose sight of the literary work, the aesthetic object and our experience of it, while examining the literary documents, just as it is important not to skim unheedingly over and through the literary documents on our way to the literary work.

A number of generalizations have been made about the relation between documents and works, but simplifications grasp one truth at the expense of others. Close observers abandon simplifications. Just two common examples will clarify. If we hold that each document is the work because without the document we would not have the work, we would have to see each different document as a different work. If one does not want to say that every copy of a work is a different work, then one must not say that the document and the work form an inseparable unity. If, on the other hand, one says a single work is represented differently by the variant texts in different documents, it seems necessary to also hold that one cannot apprehend the work as a whole without somehow holding its variant iterations in mind. Textual complexities resist simplification. How one conceives of the relationship between documents and works influences one’s practice when editing; it is important to have a sense of the complexity of that relationship.²

A third distinction is that between methods and goals. My first encounters with European, particularly German Historical Critical editing confused me because the general tendency of European editorial practice was different from that in which I had been trained. Each side thought we wanted the same goal, so, of course, the other side’s methods must be wrong. Turns out the goals were different, too. The European model emphasized assembling a record of a work’s historical forms, providing an orderly representation of textual history by combining a clear text, accurately representing one historical document’s text, with an apparatus that codified accurately the texts found in other historical documents. By contrast, Anglo-American editors conducted the same extensive research into the archival forms of the work, compiled the data for the historical apparatus, and attached that information to a critically edited text. Europeans accused Americans of contaminating the historical record; Americans accused Europeans of stopping before the real work of editing was begun. In their own way, each was right.

On the surface, the Historical / Critical approach appeared to Americans as narrow, rigid, and unimaginative, but it seems now fundamentally liberal and expansive because it presented its strictly factual product as a basis, not only for understanding the history of the texts of a work, but for new critical imaginative editing. By contrast, Scholarly Critical Editing in America, while on the surface seeming to be open and imaginative was a bit tyrannical, for it offered texts as accomplished, established facts, saying ‘This is the text you should use; the other

2 The editorial implications of different views of work and text are explored in Shillingsburg and Van Hulle 2015.

data is of historical interest only.' Historical / Critical Editing and Scholarly Critical Editing have legitimate places in textual criticism. Their methods are different because their goals are different. The test of a good edition is whether its methods actually fit its goal. And, also very important, though seldom mentioned in editorial scholarship, readers need to learn to use scholarly editions rather than assuming that they are all alike and work in the same way.

The distinction between original material documents and representations of them, displays itself in print in the distinction between new editions (especially scholarly editions) and facsimiles. New print scholarly editions claim to be only the lexical equivalent of original editions. Facsimile editions provide a simulacrum of the bibliographic code by imitating the physical aspects of original editions. The digital age has both muddled and clarified this distinction, first between originals and reproductions, and, second, between texts and images. All digital representations of print texts are reproductions; none is the original. We re-key texts or we scan images. It is an ignorant assumption that the symbols of a text can be represented in any font or any medium and still represent the same work without significant loss. Nevertheless, digitally, it is often the case that a retyped text is presented as if it were a sufficient representation of the historical text from which it derives. Bibliographers call a re-keyed text a new edition. It does not represent the work in the same way that the original document, and every keystroke is an opportunity for textual error. In digital representations of original documentary texts, the distinction between re-keyed (lexical) text and scanned (bibliographic) image is stark, requiring different files, emphasizing a distinction that always has been there.

For large modern digital projects I acknowledge two truths: No one person knows all of our areas of specialization, and every project requires expertise in several areas. Despite the tradition of literary and textual scholars working alone in small carrels in libraries, our projects now require team efforts. In one team model the chief of the project says to his helpers, Do this; do that. The helper may be a graduate research assistant, a secretary or clerk, a librarian, or a computer technologist or programmer. Like a great secret, the chief sits in the middle and knows. Except that mostly the chief in the middle does not know – does not know how to do what the helpers know how to do and does not know if the helpers actually have done it the best way or an acceptable way or have just covered up not doing it at all or not well. I believe this hierarchical model of project conduct is counterproductive, limited, and I hope doomed to extinction. Another team model involves a group of persons with similar and overlapping interests who conceive of a project and lay out a system of collaboration. The tasks are various and should go to those best suited to the task. Being a good fundraiser makes you important but does not make you a chief. Other tasks focus on bibliography; materials collection; compilation of analytical data; analysis of data; elaboration of textual principles; organization of the workflow; and selection of existing software or development of improved software, involving tools, data storage and retrieval, interfaces, navigation, and exit and portability strategies. When collaborative relationships are that of partners in an enterprise, the project becomes not only the fulfilment of an individual's concept but it can develop in ways the chief initiator

did not know and could not imagine. Of course you should vet your partners, just as they should vet you. Partnerships like chains are only as strong as their weakest link. You should strive to be that weak link – which is to say, you should strive to partner only with those who are better than yourself. It is amazing how good that will make you look.

In a moment of weakness, and although he should know better, Dirk asked me to look into the future. In my own future I see fishing, woodworking, travel for pleasure, and the superintendence of a growing array of grandchildren. In your future I see your tasks and your accomplishments through a lens that reveals that knowledge is not knowledge if it is not verified; that in editing, the facts are all documentary. If you do not have the original documents, you cannot be absolutely sure of your facts. I see that methods of editing are not facts; instead they are ways to organize and present facts. I see that editions are arguments about the facts and are susceptible to the same faults and shortcomings that attend all critical arguments. I see that you will be tempted from time to time to believe that your discoveries, your methods, and your arguments are the best in the world and that you should tell others what to do and how to do it because, like the secret, you are sitting in the middle and know.

But let me leave you with a quotation from Richard Flanagan's magical novel called *Gould's Book of Fish* (p.406), in which the protagonist says:

To be frank, though I have painted all I know, it's clear that what I know is two parts of bugger-all. All that I don't know, on the other hand, is truly impressive & the library of Alexandria would be too small to contain the details of my ignorance.

In your future, cultivate productively the bits that you do know; and try to understand what others do before dismissing them or criticizing them. The world is a big place with room for many truths, but is too small I think for error, for unsupported argument, and for attempts to make everyone see and do the same thing the same way. We just do not know enough.

References

- Dickinson, Emily. 1999. 'There is no frigate like a book (1286).' In *The Poems of Emily Dickenson*, edited by R. W. Franklin. Harvard University Press.
- Flanagan, Richard. 2016. *Gould's Book of Fish*. London: Vintage.
- Keats, John. 1919. 'On first looking into Chapman's Homer.' In *The Oxford Book of English Verse: 1250-1900*, edited by Arthur Quiller-Couch. Oxford: Clarendon.
- Shillingsburg, Peter and Dirk Van Hulle. 2015. 'Orientations to Text Revisited.' *Studies in References* 59: 27-44.

Blind spots of digital editions

The case of huge text corpora in philosophy,
theology and the history of sciences

Andreas Speer¹

Paper presented at the DiXiT Convention 'Academia, Cultural Heritage, Society', Cologne, March 14-18, 2016.

The aim of this paper

The present paper reflects on my observations inside and outside DiXiT during the past years and addresses what puzzles me about it. Outside DiXiT means, for instance, my research institute, the Thomas-Institut. Its research is mainly focused on the history of philosophy and the sciences including theology in the millennium that we usually call Middle Ages. I will not enter into the discussion on how misleading this periodization is – this is a different story and part of another debate.

My point of concern are the editions of huge text corpora which represent the scientific discourses of this millennium. Many of those editions are long-term projects and continue over generations of editors. Many of those editions – I will give some examples later – shaped the field of research, provided us with new sources for reconstructing the most influential intellectual debates and brought forth highly innovative philological and hermeneutical methods. All the more puzzling is the fact that only few of these high-ranking scholarly editions hitherto have implemented digital methods, let alone been conceived fully as digital editions. Vice versa, however, these editions also seem to be blind spots of digital editing. They seem to be out of the picture when it comes to methodological discussions and the search for technical solutions in the digital humanities. What is the reason for this puzzling state of affair?

1 andreas.speer@uni-koeln.de.

At least one reason lies in the history of digital editions in the field of digital humanities, which originated mainly from material philology, diplomatics and new philology. Therefore, digital methods focus on the most comprehensive encoding of every single detail of a witness, *e.g.* a manuscript, independent of its heuristic value for the editor. This might be worthwhile in dealing with a unique document like a charter or the manuscript of an author that went through various phases of correction and revision. But this model does not fit without serious problems those editions that I am going to talk about in this paper. These problems have to be taken into account seriously in order to bridge the gap between editions of huge scientific text corpora and scholarly digital editing.

In the following, I will ask what makes such ‘scientific’ editions special and what the overall problems with digital editions of ‘scientific’ texts are. Four case studies will show the complexity of those editions and their high methodological standards developed over the last decades, which seem to be difficult to meet in digital modeling. Finally, I will ask what could be done to enlighten this blind spot of digital editions.

What makes ‘scientific’ text corpora special?

What do I mean by speaking of ‘scientific’ texts? To answer this question, we have to go back to the invention of philosophy among the Pre-Socratics and then to the times of Plato and Aristotle, when ‘episteme’ became the technical term for a certain type of acquiring knowledge following strict methodological rules of demonstration. An exemplary model for the scope and the diversity of ‘episteme’ is defined by the writings of Aristotle, the *Corpus Aristotelicum*, which became a main subject of huge commentary work. This commentary tradition started with the ancient commentators of the Peripatetic and Neoplatonic schools, followed by the Arabic, Hebrew and finally Latin commentators, the latter being at work during the time when the European universities were founded. The *longue durée* of the Aristotelian corpus and its huge multilingual commentaries in Greek, Arabic, Hebrew and Latin, which comprise writings on logic and hermeneutics, on physics and natural philosophy in general, on ethics and politics and finally on first philosophy or metaphysics, lasted through the entire Middle Ages, the Renaissance until Early Modernity and even beyond. But what is called philosophy or ‘scientia’ comprises also cosmology and astronomy, medicine, law and theology. Speaking of theology, we have to take into consideration that theology – the name for the highest of the theoretical sciences (besides physics and mathematics) – was brought to bear – first by Christianity, then by Islam and Judaism – on the mysteries and doctrines of the respective faith, which became the cognitive content of this theological science. The most striking example is the most successful and influential source book on Christian dogmatics, composed in the 12th century by Peter Lombard, which became the subject of a huge commentary tradition over

centuries. The continuously growing catalogue nowadays comprises about 24,000 entries.²

Let us summarize some of the main characteristics of scientific text corpora:

1. They are huge in size and with regard to their transmission. As our case studies will show, one commentary can comprise more than 200 folia pages in about 100 copies from different libraries and scriptoria over a period of three centuries. We have, in fact, a 'smart big data' problem.
2. The commentary traditions are multilingual. We have to deal with translations *e.g.* from Greek into Syriac into Arabic into Hebrew into Latin. Not only do the languages differ, but so does the way and the direction of writing, which creates significant problems.
3. The scientific discourse is based on a stable terminology. Translations of scientific texts are seeking terminological stability for the sake of the scientific argument. From this intention follows a different attitude towards variation, which has to be excluded if it is not in accordance with the main goal of the scientific discourse.
4. Scientific texts address a specific readership: scholars who are interested primarily in the philosophical arguments and only accidentally in historical or philological questions. Those readers demand reliable, readable and easily accessible texts, which also should comply with the scientific *lingua franca*.

What are the overall problems with regard to digital editions of 'scientific' texts?

With respect to what I have already said, I can be very brief now. What are the overall problems concerning digital editions of scientific texts?

- the size of a text / a text corpus, which makes it impossible to encode every single detail of a manuscript in the most comprehensive way;
- the complexity of the transmission, which necessarily requires the selection of data on the basis of established philological methods;
- the specific goal of the edition, namely making a text accessible in the most multi-faceted manner for an interdisciplinary community of scholars, each member of which has a very specific, often limited interest in the respective text;
- the limited resources concerning time, manpower and not least budget;
- the missing digital tools for a conceptual as well as pragmatic handling of huge and complex semantic data.

We have to be aware that critical editions are not an aim in itself; they are part of an interdisciplinary context of rather diverse scholarly interests. Before we move on to the four case studies, which will give a glimpse of the state of the art and of the complexity of critical editions of scientific texts, let us briefly reflect

2 <http://www.repbib.uni-trier.de/cgi-bin/rebidasrb.tcl>.

on the question who uses those editions, since the editor is, to a certain extent, depending on the user of the editions due to the interdisciplinary logic of the scholarly discourse and the required support for preparing editions. In fact, there are many levels of using a critical edition. With respect to scientific text and text corpora, we can identify three main readers:

- the philosopher, who is just interested in the argument, not in orthography, punctuation, etc.;
- the philologist, who investigates the linguistic and grammatical specificities;
- the historian, who is interested in the many layers of transmission and context.

Even if this schema seems a bit oversimplified, we should think about our readers. Because it is the reader whose support the editor needs, and it is the reader who pays for the critical edition of a scientific text. For the most part, this is the philosopher. So, as an editor, you have to gain his / her support!

Four case studies of huge and complex critical editions of scientific text corpora

It has become a questionable habit at conferences concerned with digital editing to accuse printed editions of being inappropriate and lacking the methodological sophistication of digital editions. Apart from the fact that this statement does not withstand the reality check, it is neither a good starting point for one's own enterprise nor does it lead to the required cooperation with the editors, who are well acquainted with the material and who should be convinced to become engaged in digital editing.

The four case studies exemplarily show the high standards of critical editions of scientific texts, which have solved unsolvable problems and also have found models for encoding these problems. If you are able and willing to read the introduction, to decipher the apparatus and to deal with a lexicon and an index, you have the maximum of information about a complex text at hand. Moreover, those information are modeled in a way that one can deal with their complexity. Digital editions have to meet those standards and should not fall short of them. This is a great challenge, as the following cases will show. For each of those cases I am going to highlight one specific problem.

Number and Complexity: The Aristoteles Latinus, Metaphysica

Our first case is the critical edition of the Latin translation of Aristotle's *Metaphysics*, which was drafted between the second quarter of the 12th and the third quarter of the 13th century and comprises four versions translated from Greek into Latin and one version translated from the Arabic into Latin. I will particularly point to the four translations from Greek into Latin, which have been the subject of research for around a quarter of a century for Gudrun Vuillemin-Diem. She reconstructed the complex history of the translation process, which began in Venice with the so-called 'Translatio Vetustissima' by Iacobus Veneticus Graecus and was completed in Paris, where the Aristotelian *Metaphysics* became a center piece of what is commonly called 'Aristotelesrezeption', i.e. the full reception of the entire *Corpus*

Aristotelicum, which originated from the translation of his oeuvre into Latin, together with the commentaries of the Arabic commentators, especially Avicenna (Ibn Sina) and Averroes (Ibn Rušd).

The final version was produced by William of Moerbeke between 1260 and 1270, a Dominican confrere of Albert the Great and Thomas Aquinas and a philological genius. His new translation, the ‘Recensio et Translatio Guillelmi’, which originally started as a revision of the former ‘Translatio Anonyma sive Media’, is transmitted in more than 200 manuscripts plus 10 fragments, 2 manuscripts of which are lost; furthermore, there are 27 early printed editions from the 15th and 16th century. The average size of the full text is between 80 and 110 folia. A special role in this huge number of transmissions play two Parisian manuscripts (one of which is lost), which served as ‘exemplaria’ in the scriptoria of the Sorbonne. Those *exemplaria* were divided into 23 ‘peciae’, i.e. pieces, which were copied, e.g., for the classes at the university. I will come back to this *pecia* system in my next case. From the *peciae*, the editor was able to identify also the lost *exemplar* and the place of the two *exemplaria* within the *stemma codicum*, which is nothing but an attempt to reconstruct the entire history of the transmission of one of the most important translated texts from between the 13th and the 16th century. Moreover, the editor has identified the Greek codex from the 9th century (which is now in Vienna), which William was using when he first began revising the former translation and then moved to a completely new translation, which is handed down in two redactions.

PETIA 4															
In locis infra scriptis quattuordecim codices a petia alterius ordinis descripti variam lectionem praebent, quae a lectione recensionis italicae codicumque sedecim a petia prioris ordinis descriptorum discrepat :															
Liber I		Ao	Bg ³	Bx	Er	F	Kr	P ⁶	P ⁷	Pd	V ²	V ³	V ¹⁰ W	Lo ¹	Alii
15.	6 :	nullum]	secundum	+	+	+	+	+	+	+	+	+	+	+	
15.	7 :	vel]	ut	+	+	+	+	+	+	+	+	+	+	+	
15.	11 :	et]	om.	+	+	+	+	+	+	+	+	+	+	+	Er ² O ¹ P ¹
15.	22 :	eum]	ante dicet	+	+	+	+	+	+	+	+	+	+	+	
16.	33 :	quod]	si add.	+	+	+	+	+	+	+	+	+	+	+	
16.	39 :	invenitur]	invenit	+	+	+	+	+	+	+	+	+	+	+	
16.	98 :	secundum]	om.	+	+	+	+	+	+	+	+	+	+	+	C ¹
16.	105 :	igitur]	qui	+	+	+	+	+	+	+	+	+	+	+	
16.	110 :	factis]	perfectis	+	+	+	+	+	+	+	+	+	+	+	
16.	167 :	propter]	per	+	+	+	+	+	+	+	+	+	+	+	(O)
16.	182 :	est]	om.	+	+	*	+	+	*	+	+	+	+	+	As
16.	208 :	suam]	om.	+	+	+	+	+	+	+	+	+	+	+	
17.	50 :	est]	om.	+	+	+	+	+	+	+	+	+	+	+	
17.	53 :	quantum]	om.	+	+	+	+	+	+	+	+	+	+	+	
17.	62 :	modo]	om.	+	+	+	+	+	+	+	+	+	+	+	*
17.	92-93 :	ipsius... condicionem]	om.	+	+	+	+	+	+	+	+	+	+	+	
17.	105 :	vel]	in	+	+	+	+	+	+	+	+	+	+	+	
18.	12 :	aliquis laudatur]	inv.	+	+	+	+	+	+	+	+	*	+	+	Bg ¹
18.	12 :	solum]	solo	+	+	+	+	+	+	*	+	+	+	+	+
18.	121 :	inducit]	et add.	+	+	+			+	+	+	+	+	+	
19.	9 :	quae oportet]	quaecumque	+	+	+		+	+	+	+	+	+	+	(Er ²)

Figure 1: Collation table for petia 4 (Editio Leonina XLVII, 1: 91*).

The pecia-transmission: Thomas Aquinas, Sententia Libri Ethicorum

The very fact that manuscripts were copied and handed down in ‘peciae’ (oder ‘petiae’) was already well known. However, no one had thought that such a complex system ever could be reconstructed and made part of the reconstruction of the transmission of a text and its critical edition, based on the full history of the respective text. It was René-Antoine Gauthier who brought to bear the method of fully reconstructing the *pecia* system to the commentary on the *Nicomachean Ethics* by Thomas Aquinas, who composed this masterpiece while working on the second book of his most famous *Summa theologiae* in Paris in 1271/2. The editor could identify two main recensions: the *recensio Italica* and the *recensio Parisiaca*, the latter originating from two *exemplaria* from which 38 *peciae* arose. Those pieces, into which the exemplar was divided, form specific groups, which Gauthier reconstructed with the help of collation tables for each *pecia*, which show the specific transmission and grouping of the *peciae* within the transmission of the manuscripts.

Thomas’ *Ethics Commentary* comprises between 90 and 150 folia on average, and is handed down in 126 manuscripts from the 13th to the 15th century of different origin. At least 50 codices are lost. In addition, there are seven medieval editions of summaries or abbreviations, which are part of the text history, as well as 20 early printed editions.

Versions and revisions: Durandus of St. Pourçain, Commentary on the Sentences

The third case is concerned with a highly prominent example of the scientific theological discourse from the beginning of the 14th century at the University of Paris, which originated from the academic reading of Peter Lombard’s *Sentences*.³ Durand’s commentary on the *Sentences* is huge. It comprises – if one takes the early Venetian print from 1571 as a standard – 423 closely printed folio pages in double columns, which amount to more than 4000 printed standard pages without apparatus. Moreover, the case of this commentary is highly complicated because there are three main version: a first version (A) from 1307/8, which became the subject of heavy criticism so that Durandus began to revise his commentary, which was read orally to the students around 1310/1 (and which is the B version). After being appointed to lecture at the papal curia in Avignon in 1313 and becoming Bishop in 1317, he again revised his *Sentence commentary* and established this third version (C) as his standard version. An edition project at the Thomas-Institut is concerned with the two early Parisian versions, which are lively witnesses of the debates that went on at the University of Paris. The transmission, again, is highly complex: all four books of Durand’s *Sentences* commentary provide us with a completely different setting concerning the delivery and the transmission of the text handed down in 7 to 12 manuscripts each from the 14th and 15th century. Particularly Book II provides us with the clearest evidence with regard to the two redactions A and B, and gives us a glimpse into the editorial workshop of Durandus himself. Moreover, as the editor Fiorella Retucci has demonstrated, Book II shows

3 Project homepage: <http://www.thomasinstitut.uni-koeln.de/11755.html>.

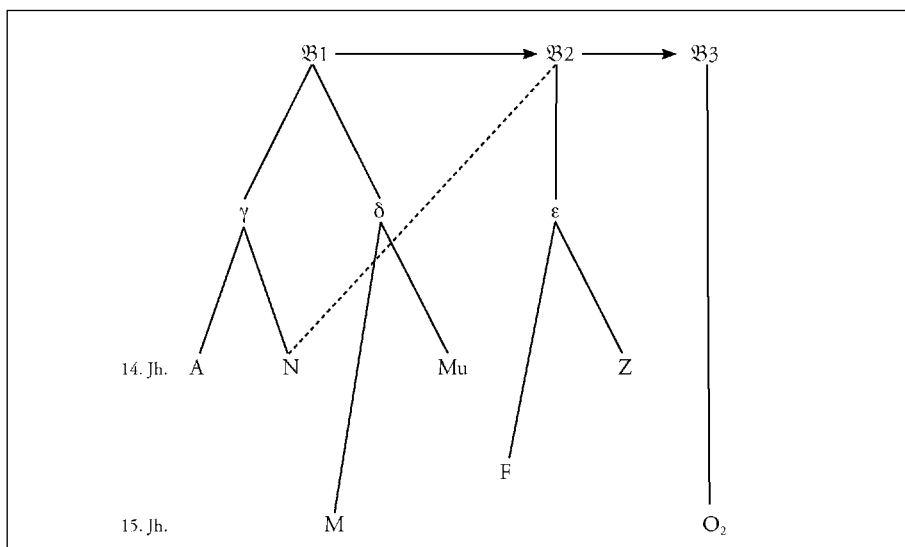


Figure 2: The 'open' stemma of Book II (ed. Retucci, 60*).

more than one redaction phase, it rather reveals a complex multistage redaction process, which contains at least three, partly even four phases.

If we assume a continuous work in the copy by the very same author, and if the changes in the text are caused by deletions and additions in the margins or between the lines in the main copy, then we have to admit an open tradition of various versions by the author himself instead of copies by one or many copyists taken from one autograph. The autograph rather must be viewed as an author's copy, on which the author was working continuously, on which he made his changes and deletions and additions. This situation ultimately leads to different phases of the transmission of the text, because it was copied – authorized or unauthorized (as Durandus complains) – at different occasions. So, individual copies could vary and might not contain the actual text of their exemplar. The critical edition, which covers and maps all the variants, therefore is built on an open stemma.

Multilingual hermeneutical levels: the Averrois Opera omnia

The commentaries on Aristotle by Ibn Rušd or Averroes (1126-1198) are a summa of the late ancient and Arabic reception of Greek philosophy. These works – and above all their Latin and Hebrew translations – have had continuous influence on central aspects in the scientific discourses for centuries. Averroes' commentaries on Aristotle's works represent a common heritage of the Arabic, Hebrew and Latin traditions of Europe, which is unmatched, both in scope and influence, by the work of any other thinker. Currently the Averroes database contains 65 Arabic, 66 Hebrew and 94 Latin works, in sum 225 works, mainly commentaries of different styles and size.⁴

Starting point were the Middle Commentaries to the so-called *Logica vetus*. The three commentaries to *Peri Hermeneias*, the *Categories* and to Porphyry's *Isagoge*,

⁴ <http://dare.uni-koeln.de/>.

ascribed to Wilhelm de Luna, are milestones in text editing, due to the efforts of the editor Roland Hissette, because they have established new methods in dealing with editions of translations from Arabic via / or Hebrew into Latin. The editions record the Latin transmission of the text against the background of their Arabic original, including the entire history of its Latin transmission, complete with the Renaissance prints and the interplay with the Hebrew transmission. In establishing an Arabic-Latin and Latin-Arabic glossary, a systematic inquiry into the translation techniques was made possible.

A new long-term project at the Thomas-Institut is concerned with the Arabic, Hebrew and Latin Reception of Aristotelian Natural Philosophy by Averroes (Ibn Rušd).⁵ This project aims at examining a central and still unstudied part of Ibn Rušd's philosophy of nature and exploring the unique interplay of the three linguistic branches of its textual transmission. The manifold interconnections within the Arabic, Hebrew and Latin transmission of the texts are reflected by the trilingual structure of the project. The project as a whole will complete the critical edition of Ibn Rušd's works on natural philosophy in Arabic, Hebrew and Latin.

To ensure a steady workflow, the *Digital Averroes Research Environment* (DARE), a platform already established at the Thomas-Institut, will be used to link and to compare the three language traditions, to integrate the existing editions and previous achievements, to secure the continuous digital analysis of the relevant materials, and ultimately to document the final as well as interim results of the project. The research platform provides tools for connecting and comparing the three language traditions and will support the editors in their work at every stage of their projects. In addition to the conventional medium of a printed edition, the platform allows for saving and publishing the results in the form of digital editions with the help of new digital research tools. Furthermore, interim results can be shared with the scholarly community early on, so that even complex long-time projects can benefit from the scientific exchange with international experts of the field. That way, DARE will continue to be a pilot enterprise in the field of interdisciplinary digital humanities by producing and distributing innovative technical solutions.

What has to be done?

The four cases clearly show how sophisticated critical editing of complex scientific text corpora has become. We have to admit that those text corpora are a true challenge for both traditional and digital philology with respect to the level of complexity regarding the transmission of texts, the diversity of the context and the methodological efforts. All those information can be found in the printed editions. What is then the surplus value of a digital edition? Why should editors of such complex projects add complexity to their already hypercomplex undertaking? These questions have to be addressed seriously, because they are crucial for the future of digital editing, in particular if one makes the claim that future scholarly

5 <http://averroes.uni-koeln.de/>.

editing will be and has to be digital. So, what are the conditions on which this might happen? What has to be done?

1. Digital editing has to meet the philological standards of scholarly editing. Technical encoding standards like TEI / XML are not philological standards and they do not add much to the philological and conceptual design of a project if they are not integrated pragmatically into a workflow that gives priority to the philological work.
2. If digital editing should become the standard practice for preparing editions, digital tools, which are easy to handle and do not require much technical or even programming skills, are needed. Moreover, we need useful standardization processes, which lead to an unhindered and unrestricted usage of digital tools.
3. Caring for research data has gained growing importance, also in the humanities. This is particularly true for critical editions, which are based on a mountain of material evidence, as our four cases have shown. Databases and research platforms have become important tools to collect and to share research data and to create international networks to cope with the new challenges. Among long-term projects of complex scientific editions, there are already examples for such databases and research platforms.⁶
4. Digital editions facilitate a modular approach by a wide range of digital tools and methods, which will provide the means for the implementation of existing editions and preliminary studies as well as for the documentation of the final and interim results. The release of interim solutions and intermediary steps of a complex 'work in progress' fosters scholarly discussions and the exchange of ideas.
5. We should overcome the opposition between the printed and the digital edition. Digital editing still has to demonstrate its strength in solving complex problems in a sustainable manner. Therefore, both ways of editing should coincide. This is particularly true for complex and huge scientific editions, which often are long-lasting projects with a long history behind and ahead of them.
6. If we want to foster digital editing and to make it the standard among critical editions of complex scientific texts and text corpora, we need realistic concepts for a project design which matches the requirements and the specific research interests of those editorial projects. The strategy of doing small parts in the most ideal manner and leaving the rest for a future which will probably never come is not the winning strategy!
7. The key for all those problems is: Doing science! That means that the problems should not be applied from outside and from a merely abstract point of view, but rather from within a concrete project and due to its specific research interest and project design. The proof of the pudding is in the eating!

⁶ See e.g. <http://ptolemaeus.badw.de>, <http://www.averroes.uni-koeln.de/>, <http://dare.uni-koeln.de>.

References

- Aristoteles latinus. 1995. *Metaphysica*, lib. I-XIV. *Recensio et Translatio* Guillelmo de Moerbeka (vol. XXV 3. 1-2), (ed.) G. Vuillemin-Diem. Leiden, New York, Köln: Brill.
- Averroes Latinus. 2010. *Commentum medium super libro Praedicamentorum Aristotelis*. *Translatio* Wilhelmo de Luna adscripta (vol. XI), edited by R. Hissette. *Apparatu arabo-latino supplementoque adnotationum* instruxit A. Bertolacci. *Lexica confecerunt* R. Hissette et A. Bertolacci. *Commendatione auxit* L. J. Bataillon (†). Lovanii: Peeters.
- Durandi de Sancto Porciano *Scriptum super IV Libros Sententiarum Distinctiones* 1-5 libri secundi. 2012., edited by F. Retucci, RTPM-Bibliotheca 10. 2. 1. Leuven, Paris, Walpole: Peeters.
- Sancti Thomae de Aquino *Sententia Libri Ethicorum*. 1969, edited by R. -A. Gauthier, 2 vols, *Editio Leonina* tom. XLVII, 1-2. Romae: Ad Sanctae Sabinae.
- Speer, A., F. Retucci, T. Jeschke and G. Guldentops (eds). 2014. *Durand of Saint-Pourçain and His Sentences Commentary. Historical, Philosophical, and Theological Issues*. Leuven, Paris, Walpole: Peeters.

Data driven editing: materials, product and analysis

Linda Spinazzè,¹ Richard Hadden²

& Misha Broughton³

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

The remediation of cultural heritage documents into a digital environment – particularly through the disparate but related practices of mass digitization and digital scholarly editing – has a keen focus on textual and multi-media content. However, this focus sometimes occludes the fact that, working within a digital workflow, our core material is, in fact, data. This panel seeks to explore the possibilities of a more data-driven editing practice, one that sees not only our material (digital proxies, collections information, transcriptions, and metadata) but also our resulting products (corpora, editions) and all of our intermediary stages not as text or images or content, but as data per se. In the following sections, we will seek to reconcile the ambiguity inherent to humanities inquiry with the exactitude required of digital data, asking how we can 'read' this data, and what – if anything – is our responsibility as editors to provide access not merely to the final argument of our editions, but to the data that informs it.

1 linda.spinazze@gmail.com.

2 richard.hadden@nuim.ie.

3 wbrought@uni-koeln.de.

Source material as data – Linda Spinazzè

The case study which follows is concerned with crowdsourcing digital editing. In fact an overview of *Letters of 1916* project provides an occasion to explore the particular way in which a collection of texts can be edited digitally according to Web 2.0 philosophy of sharing and collaboration.

The *Letters of 1916* is a work in progress to create an online fully searchable collection of correspondence. The aim of the project is to gather and edit letters in the period leading up and just after the 1916 Rising by engaging the ‘crowd’. The collection includes private letters, business missives, official documents, postcards, greeting cards and telegrams written around the time of the Easter Rising of 1916. On 24 April 1916, Easter Monday, in Dublin a small group of Irish nationalists decided to rebel against British rule. The General Post Office (GPO) served as the headquarters, where seven members of the Council who planned the Rising declared the proclamation of Irish Republic in front of this building. Within a week, the British army quickly had suppressed the rebellion and, on 3 May it started to execute the leaders of the Rising. Even though Ireland did not gain independence until 1922, it is a common opinion that the Easter Insurrection is the moment when everything changed, it is considered by historians as a sort of ‘point of no return’⁴. The *Letters of 1916* project aims to help in understanding this ‘change’ better creating the new collection consisting of pieces of correspondence written between the 1st of November 1915 and the 31st of October 1916. Assuming that the words present in the letters⁵ are the witnesses of different aspects of the society in that particular historical period, we are aware that such a collection can open new perspectives on the events and daily life at that time.

In contrast to a more ‘traditional digital collection’ which tends to be linked to a physical archive stored in a library, or for example, to a specific author already studied and edited, the *Letters of 1916* project brings together images of the correspondence from many different institutions, about 20 and also from private collections⁶. So, not only is the team or experts responsible for the upload, but often members of the public⁷ undertake the process of uploading their family letters from scratch thanks to the platform created by the *Letters of 1916* team. In terms of crowdsourcing, the native platform of the project utilizes the Omeka software⁸ alongside some plugins which carry out specific functionality.

4 The bibliography is huge; for a general reference about the subject we can just refer to one of the most recent (McGarry 2010) (see also the new ‘Centenary edition’, published in 2016).

5 It is worth to point out to this fact: «In 1914-1915, the last fiscal year during which records of letters posted were kept, approximately 192 million letters were mailed within Ireland, which works out at roughly forty-four letters per person», in Novick 1999: 350.

6 Here the list of institutions which have allowed us to include images of letters and photographs in the *Letters of 1916* project: <http://letters1916.maynoothuniversity.ie/learn/index.php/collaborate/institutions>. Accessed 6.10.2017.

7 For a critical perspective on the gap between crowdsourcing and mission and values of cultural heritage organisations see Ridge 2014.

8 The *Letters of 1916* uses the Omeka 1. 5. 3 (<http://omeka.org/>); the transcription interface is based on the *Scripto* plugin <http://scripto.org/>; see forward for other add-ons. Sites accessed 6.10.2017.

Because of the crowdsourcing nature of the project is particular interested in the large participation of *amateurs*⁹ and in the creation of a ‘corpus that never was’¹⁰, in such a digital collection the contact with the ‘original text-bearing’ objects is particularly fleeting. Precisely because the workflow quickly moves away from the material objects in favour of the digital data, the conversion from the ‘material’ to the ‘digital’ has to be particularly accurate. After taking the high resolution images the user has to upload the digital item via a form which helps to simultaneously create some basic metadata (such as title, creator, place¹¹). In filling in this form, it should be clear, especially to the non-specialist that in this first phase they are contributing in creating a basic digital storage, that it is not a plain silo of photographs, but an actual database of items – of actual structured digital items. The high resolution digital images are surrogates of the original letters, and the archival of this digital material guarantees its curation and preservation. This is especially true in the case of certain private collections which often are stored inappropriately in their physical form (see Figure 1). After the uploading and structuring of the metadata, the new items are quickly revised by a member of the team who makes them accessible in the transcription area of the site, just ready for the next phase of the *Letters of 1916* workflow of editing.



Figure 1: Sometimes the private collections are not stored in the appropriate way, other times the letters are hidden away for decades. Here is an example of a metal biscuit tin filled with old letters and found in an attic by one of the contributors of the Letters of 1916 collection. Photo: ©Letters of 1916 (Kildare Launch, Maynooth University, May 2014).

9 For a definition of *amateur* inside the crowdsourcing philosophy, Owens 2014.

10 Paraphrasing the ‘text that never was’; see Greetham 1999, or more recently 2014.

11 See form at: <http://letters1916.maynoothuniversity.ie/images/HowToUploadALetter.pdf>. Accessed 6.10.2017.

In order to also provide the community of users from the public audience with an introduction to TEI mark-up and more generally to a digital scholarly editing workflow, at this point the *Letters of 1916* project does not require a simple transcription but rather a 'structured' one, which incorporates basic encoding too. To combine the requirement of a TEI structured transcribed text the plain text-field by Omeka/Scripto is equipped with an adapted version of the 'Bentham toolbar'¹². This plugin serves as a method for encoding some main feature contained in the letters (the features which provide information about the material aspect or 'semantic' details¹³).

In order to ensure that a digital scholarly edition is created from all these transcriptions, the team editors have to handle encoded information with formal errors, misunderstandings, omissions. In fact, inside the definition itself of 'edition', the accuracy assessment is one of the basic requirements. The question of how to dynamically proof the accuracy of the tagging remains. When the error is not about the 'well-formedness' of the mark-up, but is a real misinterpretation of the tag or a completely wrong reading, it is almost unpredictable. Is there a dynamical solution?

At the moment, we are concerned with figuring out a solution for proofing this kind of collection 'driven' by digital data on its own is unrealistic. So, considering that the human checking is necessary, the question is: how can we combine automated and manual editing effectively? And more importantly, can we just consider an edition a plain transcription, even if it is well structured and well formed?

The edition as data – Misha Broughton

If the resources of digital text editing are data, it is important to also note that its output is equally data. While this may seem self-evident in theory, it is a point easily forgotten in practice, where the aim of editing is so commonly the production of an edition. However, while this goal is certainly natural, our concept of what an edition is, or can be, is still limited to a print concept, or as Patrick Sahle would have it, 'the print paradigm.'¹⁴

What is the nature of a critical edition, print or otherwise? The *MLA Guidelines for Editors of Scholarly Editions* states that its 'basic task is to present a reliable text,'¹⁵ with – I argue – a silent emphasis on the singular 'a.' Editions are composed, however, from multiple document witnesses and, often, a contentious transmission history. If this is the case, then editing – or, at least, editing to the edition – is a process of ablation, of whittling away at textual extraneities that do not support the privileged reading of that particular edition. And yet the text, the holographic, syncretic whole that we aim to represent through our endeavors, is surely bigger

12 TEIToolbar from Transcribe Bentham project: <http://www.ucl.ac.uk/transcribe-bentham/>. Accessed 6.10.2017.

13 See explanation at: <http://letters1916.maynoothuniversity.ie/images/ProofingXMLGuidelines.pdf>. Accessed 6.10.2017.

14 Sahle, Patrick. 'about'. *A catalog of Digital Scholarly editions, v 3.0, snapshot 2008*. Accessed 6.10.2017.

15 'Guidelines for Editors of Scholarly Editions'. Modern Language Association. Accessed 16.1.2016

than any single reading, just as it is bigger than any single document witness that attests it. And yet, in effect, this is what the edition conceived under such these terms can not help but be: another document witness in the text's transmission history, albeit one authorized by an expert scholarly editor.

In a previous technological environment, the edition *could* be nothing but. Limited by the same constraints of the page space in which previous document witnesses were compiled, the print scholarly edition had but few methods (e.g. the critical apparatus, paratext, footnotes, marginalia) to do anything besides document the textual history largely as it was received. Though the advent of digital media technologies has brought many new affordances to the display, publication, and discoverability of scholarly editions, it brought little – if any – reconsideration of what the edition is. In our practical commonplaces, like the MLA Guideline cited above, we have re-inscribed the familiar shape of the print-document edition in the digital: a single, reliable text with a scattering of apparati to record the more important variation. Perhaps more importantly, though, this understanding of the edition as a print-like document also has influenced the logical model which informs our most prominent data model, TEI/XML. The Ordered Hierarchy of Content Object model of text,¹⁶ which informs the XML markup language, was proposed as a method of organizing text data *specifically* for its similarity to print documents. For as far as we have come, technologically, we have arrived at little more than print documents migrated whole-cloth from pages to screens.

It is important to remember, though, that while the document-like (or text-like, if you will) methods of organizing data are a very venerable and mature technology, they are still only one possible method, and one far from perfect for all applications. While the form is familiar to editors for its similarity to the witnesses it collates, it is this very similarity that limits it dimensionally, making compositing of various textual features difficult, at best. And while it is certainly needful for the presentation of the 'reliable (reading) text' aforementioned, considerations of *presentation* and of *encoding* need not (and should not) be confused.¹⁷

These concerns would be entirely academic, of course, if not for the fact that the practice already is running afoul of all-too practical consequences of this document-like approach to encoding. The problems of hierarchy overlap (Renear *et al.* 1996) and limited data interoperability (Schmidt 2014) are, I argue, not only related but both stem from the same dimensional limitations imposed on digital textual encoding by a print-centric conceptual model and encoding scheme. It is all but impossible to fully represent a topographically complex three dimensional object in a two dimensional plane. How much more difficult must it be, then, to represent the layered complexities of multiple document witnesses – each at least a two-dimensional page space and some with their own collections of dimensionally extending commentaries, emendations, and apparati – in a conceptual space utilizing only the same set of dimensions and functions? The TEI community

16 See De Rose *et al.* 1997.

17 At present, I will leave the definition of 'reading' as the rather conservative one of reading linearly page-by-page (or its screen equivalent), though discussion certainly is warranted of the relation of the DSE to Moretti's 'Distant Reading,' Bloomer's 'unruly reading,' and the large corpus of early work on hypertext reading practices.

has done wonders adapting to the shortcomings of the model, sometimes at the expense of the underlying logic of its assumptions.

For all of that, though, I predict that these problems – and more like them that we have not yet considered – will multiply far beyond our ability to make allowances for them under existing practice. Our understanding of text in its material form has been expanded by our years of work transmediating its content to the digital and it is this expanded understanding of (often printed) texts which leads to the desire to encode features or sets of features which challenge the underlying assumptions of our practice. What is needed is not more allowances in the current technologies to ‘make it do what we want,’ but back trenching, a reconsideration of the logical model by which we encode that allows native expression of the dimensional complexity we have come to understand in text. In short, we must *display* our editions, but *encode* our data. Such an encoding would fulfill the requirements of what Elena Pierazzo has called the ‘paradigmatic edition,’ an encoding that provides ‘many alternative options for the same string of text in a nonlinear way,’ (Pierazzo 2014) though perhaps going a step further to allow even different strings of text, different readings, different *editions*, in the same encoding.

However, if our encoding is not organized along the familiar and readily legible modes of the text document, how should it be organized? Even the most basic database or markup language provides a wealth of methods to interconnect related data, with features allowing for the relational or associative linking of content. My own proposal, currently under development in my doctoral dissertation at the

Text as Reproduction of Textual Objects

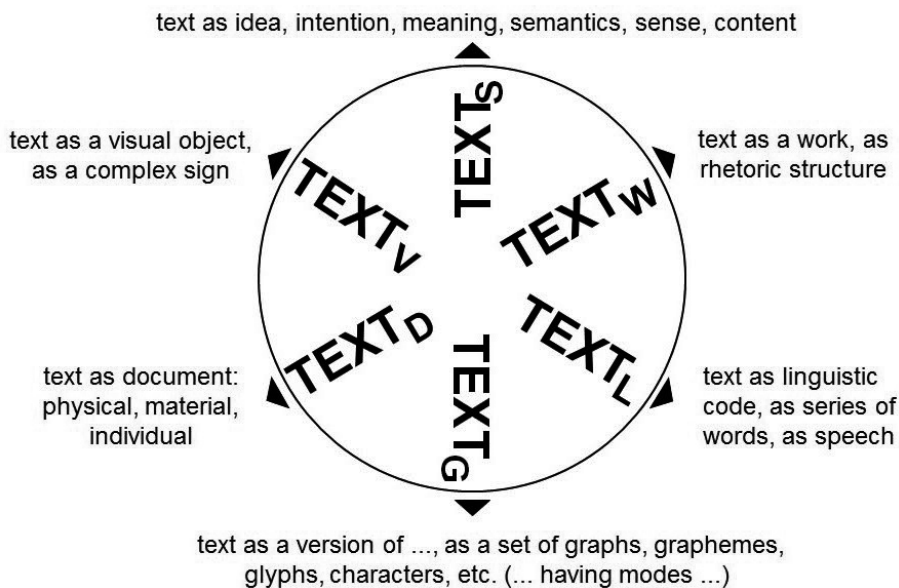


Figure 2: Patrick Sahle's Wheel of the Texts. Figure used by permission.

University of Cologne, is based on Patrick Sahle’s ‘Wheel of the Text’ (Figure 2, seeing ‘the text’ as a locus of interpretation of discrete text-bearing objects, with features and values for observation dependent on the perspective of the observer. However, while Sahle’s wheel indicates an equality of these perspectives, none alone sufficient to fully account for the features of other perspectives, my own approach sees a chain of *necessity* from the most document-centric perspective to the more text-centric perspectives. For instance, if we can not say that observations of the document-centric perspective are entirely sufficient to justify our observation of a linguistic code in the same text or the linguistic code of the work, we must say that the presence of a document is necessary to claim that a linguistic code is being employed and that the use of a linguistic code is necessary to the presence of a textual work. While our object of inquiry, then, is the abstract text, the ‘communicative act’ (Robinson) that is embodied merely in documents, we must acknowledge the presence of instantiated, embodying documents to make any claim that such an abstract text exists. Counterintuitively, perhaps, I propose that the best way to free this abstract text from the confines of a document-centric organizing mode is precisely by encoding data directly observed from documents and linking successive layers of sinterpretative perspective atop it.

The advantage of this system is three-fold: first, by separating layers of interpretative perspective, it provides a measure of vertical independence, separating the various observational perspectives from Sahle’s wheel and thus avoiding hierarchy overlap common when trying to encode such features together in an in-line transcription. Second, it provides a measure of horizontal independence, allowing for the encoding of disparate editorial perspectives or features clusters in distinct groupings without reference to other perspectives that reference the same base (see Feature A/B/C in Figure 3). Third, though not represented in Figure 3, this approach allows for the encoding of *depth*, allowing even for the encoding of contradictory or mutually exclusive interpretations from the same editorial perspective (e.g. disagreeing transcriptions of the same region, differing tagging of prosody of the same transcription, etc.).

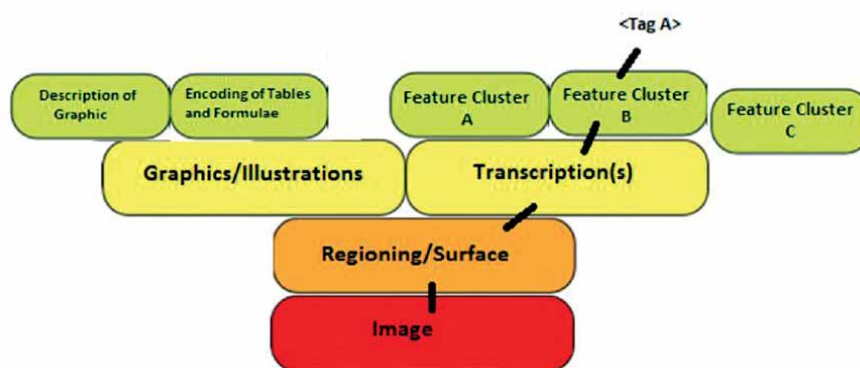


Figure 3: A layered approach to text modelling.

Editing for data – Richard Hadden

Digital Scholarly Editions have tended to follow a particular paradigm from their printed days, albeit a largely invisible one. This is the tendency to conflate the results of the actions of editing and the form of the edition. Such a view is natural enough: the text of the edition in print is bound very tightly to the material form in print, with specific adaptations designed to represent particular kinds of texts. The critical apparatus, for instance, may be seen as a way of representing the text of multiple witnesses, layered upon a base text. In this case, the presentation is coupled very strongly to the form of the edition.

It is possible – even arguably necessary – to consider a digital edition in the same light. If we view text as anything other than a pure abstraction, it is clear that the representation of text on screen is as vital to a reader's understanding as it is in print. Following Patrick Sahle's theory of a 'pluralistic' understanding of text (modelled as a 'wheel of text'), one can argue that the text of any edition – which is to say, that viewed by the reader – is the 'totality' of this plurality. If we ask, therefore, what is the edition in a digital, it is the text (as encoded) combined with its presentation.

Such a perspective, while valid, ignores the fundamental difference between a digital and a print edition: notably, that the text of the edition is stored as an abstraction, and only rendered in some form of interface on-demand by a reader's computer (or other device). As a result, there is inherently a disconnect between the edited text as an abstraction – data – and the edited text as rendered. What I will argue is that, though theoretically it is impossible to fully comprehend an edition and the text represented therein through only one aspect of its plurality, pragmatically at least we, as editors, should concentrate more forcefully on editing data, as that is what fundamentally drives an edition. To do otherwise is to ignore a fundamental reality of text in digital form, and, indeed, to deprive ourselves of a major benefit of the medium.

This is not to suggest that such a view is not already partially applicable. Since the bad old days of tags and inline styling in HTML came to an end, web development already enforces a degree of separation of style (described using CSS) and data (encoded using HTML). With the advent of HTML5 and a greater range of semantically-meaningful, rather than presentationally-oriented, elements, such a divide is even greater. This is one abstraction of the text, albeit one only applicable to a web browser. Using TEI-XML to encode texts, a standard practice, increases this divide. We are able to describe text in much less generic ways (compared to HTML5). Further processing, using XSLT, for instance, to transform XML into HTML, to which CSS then can be added, which then can be rendered by a browser.

I would argue, however, that despite these separations of concerns, there is still too great a tendency to consider TEI encoding as merely the first step towards building an edition. Even though the actual building of an edition website may be the responsibility of someone with greater expertise – i.e. a project may employ editors to encode text and a web designer to build the site – too great a focus is placed, at the stage of encoding, on the end product. That is to say, we are too ready to abandon a greater level of expression (TEI) in order to produce a website

with some text on it. Altering the focus towards editing data – as an end in and of itself – rather than editing towards a final product, seems a way to avoid what, ultimately, appears to be work for no end.

In a recent paper Peter Shillingsburg argues that producing digital editions is too complicated, compared to the days when he could edit text and typeset the final edition (using LaTeX) all by himself. Now a greater range of expertise (web design, data processing, not to mention arcane procedures of server configuration) is required (Shillingsburg 2015). This is true as far as it goes, but to build, as he suggests, a system that would take care of everything is to lose sight of the benefits of the separation of data and presentation. There is no reason *per se* that he could not encode his edition directly into HTML – after all, if one can learn LaTeX, learning another relatively simple markup language and vocabulary cannot be too difficult; the two are broadly analogous. Such an approach, however, would involve throwing away a degree of abstraction, and ultimately constrain the use (or re-use, or elaboration) of the edited text.

If we edit towards data rather than an edition, we run into at least some conceptual problems, not least: what exactly are we making? I would argue that a TEI document is, in itself, an edition, with at least equal status to a beautifully-rendered and functioning website. After all, it is (for me at least) as easy to ‘read’ a TEI-XML encoding of a text as to understand the arcane symbols employed in, say, a typical printed critical edition. At the same time, it must have a degree of primacy: a website built by transforming XML into a web-based interface is clearly derivative. As a result, it is not possible to completely disregard the end product when editing the data. Encoded data is not a neutral, ‘pure abstraction’ – as can be said of any form of editing – and neither is it total. If we wish, therefore, to produce a certain kind of edition, it is necessary that enough detail is encoded to make this possible. But we should aim for a form of neutrality – or, better put, a degree of agnosticism with regards to the final product. This is, after all, what the TEI does, by inviting us to describe the text of a work or document rather than its endpoint.

The great benefit of this is both in the re-use and further elaboration of data. Re-use is, of course, one of the fundamental points of the TEI: by providing a set vocabulary, it should be possible for the data created by one project to be reused in another (many digital scholarly editions make their TEI data available for this purpose). However, this potential is seldom realised, I would suggest chiefly because even TEI encoding is geared in too great a degree towards its end transformation into a HTML. As projects necessarily are limited in scope, this is hardly surprising: editors encode as much information as they need, and in the way that they need it, to produce the kind of edition they aim to make.

An approach to circumventing these obvious restrictions is to treat editing as a form of ‘progressive enhancement’ (to borrow a web design term) of data: editing is treated as a modular and incremental workflow, where the objective is to elaborate upon the data as it exists, so far as this might allow new ends to be achieved. Such tasks may be carried out by those working on the initial project, or (re)users of the data further down the line. Moreover, elements of data already encoded may be used algorithmically by automated processes and scripts.

Application to the *Letters of 1916* project

Linda Spinazzè already has described some of the workflow of the *Letters of 1916* project. I aim here to outline how the principles of this data-centric approach to editing have been, are being, and (I hope) will be employed.

The first aspect to note is the very clear delineation of phases in the project workflow, in terms of activity and personnel involved (this is one distinction from Shillingsburg's desire for end-to-end production of his own editions). Part of this is, of course, necessary as a result of the crowdsourcing nature of the project. The first stage is the capturing of digital images of letters en masse, by project team members visiting archives (I should say 'principally by', as some, though a small minority of images are uploaded directly by contributors). The letter images are uploaded to be transcribed by the 'crowd', who also add a limited number of TEI-XML tags (not necessarily accurately). At this stage, we have data that arguably can be distributed as an edition – albeit not a very good one.

The transcribed data then is extracted from the crowd-transcription environment (Omeka) and enhanced using a range of automated scripts written in Python. The text, which is stored as individual pages in Omeka, is joined into a single TEI document for each letter; metadata added to the letter in Omeka is used to construct TEI elements such as <correspDesc> and <revisionDesc>; and further semantic information is added automatically based on the limited encoding already completed.

The 'compiled' TEI documents then are sent to be proofed for text and markup by project team members using a purpose-built, web-based editing tool, which tracks edits to the documents using automatic commits to a git repository. At this stage, further data is added, such as normalising names of senders and receivers from a canonical list. The workflow thus far is strongly data-focused, with effort geared towards producing accurate and valid TEI encoding; also, each stage can be viewed as a progressive elaboration of semantic information over the 'base' transcription.

At this stage, it is necessary to consider the plans for the forthcoming edition. This has been designed by another project member, and is designed as a full-text searchable edition, with a provision of the full letter-text and side-by-side page and image views. This new site has been designed to store text as pre-rendered HTML. However, it uses only some of the encoded TEI elements. As such, it can be seen as consumer of the edition data, while the focus of editing remains on the data itself. The new edition's importing process is adapted to the data, rather than the other way round. By using TEI, much of this adaptation can be foreseen; though where this is not the case (for instance, the use of specific elements to indicate document structure), it is for the importing scripts to adapt. This being the case, and following the argument made thus far, it can be seen as one consumer among many potential consumers: it satisfies one potential use of the data, but by no means all of them.

As a result of such an approach, it is possible to envisage further uses for this data, both in terms of alternative editions, possibly using data-analysis techniques such as topic modelling, and, more importantly perhaps, further elaboration of the base data: thus far, encoding has steered clear of more graphical features of

the document (such as official stamps) which could be added later; the marked-up addresses can be used to add geolocation data. Moreover, data already marked up could be used to train classifiers to automate the markup of the next ‘generation’ of letters to pass through the workflow: work on this has been attempted already, using decision-tree classifiers to identify lines in the text with particular significance, such as addresses and dates. Such an approach also can be used to identify named entities within the text body, which currently are not marked up.

The obvious downside to such a data-driven approach is consistency. If the underlying data of an edition is constantly – and actively – changed, what are the implications of this for a scholarly edition, of which academic rigour demands stability? To allow versions to exist concurrently, the project uses two approaches. Firstly, the data is stored in a git repository, which tracks all changes to documents, and also allows the data to be cloned, edited and re-merged as necessary. Further to this, the TEI markup makes extensive use of the <revisionDesc> element: each change to each page made by transcribers is logged as a revision, with the text of each ‘version’ stored in an XML comment (this is necessary as the transcribed text is not necessarily valid XML) for future reference. Each scripting operation logs its effect in a revision as well.

As with the TEI-encoded text itself, this revision data is not oriented towards a particular use: instead, it is simply made available to potential consumers to make use of as required.

```
<revisionDesc>

  <change when="2014-05-25T19:26:34" who="#Badzmiek"> Page 2637 modified:
    <!-- Holy Ghost Missionary College
Kimmage Manor,
Dublin 1 July 1916

Dear Monsignor Hogan,
I cannot but accede to your request to conduct the Retreat for the opening of the coming Scholast
Please let me have a copy of the Regulations usually followed, and I should, also, like to have a
With every best wish,
Faithfully yours,
John J. Murphy C.S.S.P.
The Right Rev. J. F. Hogan, D.D.,
President, &c. -->
  </change>

  <change when="2014-06-28T15:24:21" who="#Badzmiek"> Page 2637 modified:
    <!-- <address>Holy Ghost
Missionary College.
Kimmage Manor,
Dublin</address> <date> 1 July 1916</date>

<salute>Dear Monsignor Hogan,</salute>

<p>I cannot but accede to
your request to conduct the Retreat
for the opening of the coming Scholastic
Year. So you may count on me for that
purpose.</p>
<p>Please let me have a copy of
the Regulations usually followed, and
```

Figure 4: Illustration of the revisionDesc in the Letters of 1916 TEI files.

This final point illustrates the pitfall of this data-centric approach. With each phase divorced from the next, and with a greatly lessened possibility for revision of a previous process at a later stage, rigour at each point is essential. Each elaboration of data is built upon a pre-existing foundation, which must be secure. At the same time, the benefits for ongoing usefulness of editorial activity make such an approach worthwhile.

References

- DeRose, Steven J., David G. Durand, Elli Mylonas and Allen Renear. 1997. 'What is Text, Really?' ACM SIGDOC Asterisk Journal of Computer Documentation.
- Greetham, David C. 1999. *Theories of the Text*. London: Oxford University Press.
- Greetham David C. 2014. "Retexting the Barthesian Text in Textual Studies." In *The Conversant. The Renaissance of Roland Barthes, a Special Issue*, edited by Alex Wermer-Colan. (<http://theconversant.org/?p=7880>).
- McGarry, Fearghal. 2010. *The Rising: Easter 1916*. Oxford: Oxford University Press.
- Novick, Ben. 1999. "Postal censorship in Ireland, 1914-1916." *Irish Historical Studies* 31(123): 343-356.
- Renear, Allen, Elli Mylonas, and David Durand. 1996. 'Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies.' In *Research in Humanities Computing*. Oxford University Press.
- Ridge, Mia. 2014. *Crowdsourcing our Cultural Heritage*. Farnham-Burlington: Ashgate.
- Owens, Trevor. 2014. "Making Crowdsourcing Compatible with the Missions and Values of cultural Heritage Organisations." in *Crowdsourcing our Cultural Heritage*, edited by M. Ridge. Farnham-Burlington: Farnham-Burlington: Ashgate: 269-280.
- Pierazzo, Elena. 2014. 'Digital Documentary Editions and the Others.' In *Scholarly Editing: The Annual of the Association for Documentary Editing* 35.
- Robinson, Peter. 'The Concept of the Work in the Digital Age'. Pre-publication draft.
- Schmidt, Desmond. 2014. 'Towards an Interoperable Digital Scholarly Edition'. *Journal of the Text Encoding Initiative* 7.

Making copies

Kathryn Sutherland¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Editing in the digital medium offers multiple challenges of a foundational kind: whether they be the urgent economics of the digital (getting the funding in place); the politics of the digital (the institutional and interdisciplinary collaborations required of the edition-builders – scholars, archivists, and information technologists); or the need to anticipate how the scholar-user or the interested reader-user will engage – the flexible entry points demanded of the digital edition. Interestingly, too, in Britain we currently are experiencing a significant reflowering of print editions of major works: from Cambridge University Press, the Works of Jonathan Swift; from Oxford University Press, John Donne's Sermons, Robert Burns's Poetry and Prose, and new editions planned of the poetry of Pope and Wordsworth. All have been conceived with no more than a back-end digital component, an appended digital laboratory or home for excess textual matter rather than a primary instantiation as web-based edition, while the commercial publishers Faber and Palgrave recently have entered the market with print-only scholarly editions of the poems of T.S. Eliot and the Italian Letters of Elizabeth I.

The recent major reconnection with print is not irrelevant to our current digital preoccupation with the copy, with making copies. The critical work of copying, whether as transcription, collation, or bibliographic description traditionally has been a fundamental editorial activity. In the digital environment it is supplemented and potentially overtaken by a further aspect: the reproduction of images of actual objects (whether of old editions or 'raw' sources, such as holograph manuscripts). This is worth pondering because, although facsimile reproduction within print editions has a long history, facsimile editing (until recently) rarely has contributed to the work of scholarly editing. This might be explained as the problem of visual evidence; the fact that it is evidence by illustration rather than by analysis, and that illustration has tended to smack of something belated and nostalgic – bibliophilic rather than bibliographic.

¹ kathryn.sutherland@ell.ox.ac.uk.

Largely unacknowledged in American editorial theory and practice of the 20th century, routine facsimile reproduction was confined in Britain over the same period, in the dominant Clarendon Press model, to the use of the photo-facsimile first-edition title-page, a gesture to history that sat oddly with the compressed, non-historical, eclectic edition (never a straightforward reprint) that lay beyond it. This was not because facsimile editions were necessarily poor in quality; photography was capable of great precision early on. It had more to do with the abstract and synoptic allegiances of theories and practices of scholarly representation favoured by most editors. Editing conventionally implies an engagement beyond reproduction, in which the act of copying is only one stage, albeit an essential stage, in the critical interrogation of the evidence.

Our digital editorial preoccupation with making copies cannot be put down simply to the fact that the technology allows it – that computers are good facsimile machines; something more is going on. This ‘something more’ has much to do with the radical reconception of bibliography (editing’s underlying ‘grammar of literary investigation’; Greg 1966, 83) as book history (with its sociological parameters) and with how that feeds into a preoccupation with the digital environment as a medium for replicating materialities other than its own.

In the age of the digital copy, the book, a robust and enduring reproductive technology, has gained new glamour – each copy seemingly an original with unique messages to deliver: annotations in its margins, the somatic traces of long-dead owners, and the identifiable signs of the agents of its manufacture. What was, until relatively recently, the narrow and arcane science of the bibliographer, practised in the detection of leaf cancels, irregular collations, and lying title-pages (signs of the duplicity of the seeming copy and evidence of the errors of history) has entered the mainstream of our critical thinking as something altogether warmer and more generous: book history as a kind of biblio-biography. Books, their feel and heft, their bindings, their paper, all assume profound significance and repay our minutest attention. We increasingly are impelled to curate each copy as ‘a protest against forgetting’ (Obrist 2011, 27) – forgetting what, exactly?

What has the affective relationship offered by copies of books to do with issues of editing, in either print or digital forms? For open enthusiasts like Katharine Hayles and Jerome McGann, the digital is ‘the lever for enlightenment’ upon the book. ‘Literature’, writes Hayles (and McGann echoes her) ‘was never only words’ (McGann 2013, 278-279). For them, digital media, rather than print, indeed not print under any circumstances, provide the tools that give critical purchase on literature’s ‘bibliographic interfaces’, on all that is implied in its shared and individual design features. Bring on Don McKenzie, the book historian hailed by McGann as ‘The Hero of Our Own Time’ (281), and we hear something that only appears to offer an opposing argument. Back in 1992, when major repositories like the British Library were announcing plans to remove bulky volumes from their shelves to rely at that stage on microfilm surrogates, McKenzie made an important defence of original artefacts, arguing:

Any simulation (including re-presentation in a database – a copy of a copy) is an impoverishment, a theft of evidence, a denial of more exact and immediate visual and tactile ways of knowing, a destruction of their quiddity as collaborative products under the varying historical conditions of their successive realisations.

(McKenzie 2002, 271)

On the one side, the digital provides our only purchase on literature's 'bibliographic interfaces'; on the other, the digital is 'an impoverishment, a theft of [bibliographic] evidence'. The apparent stand-off in the two positions is resolved, I would argue, by recourse to the principle that underlies the science of bibliographic description, where the challenge is to represent, in some other form than itself, an entity (in this case a particular book copy) that is necessarily absent. It is as if the unreproducible features of the book-object – its paper, binding, sewn structures, inking – come into full view only in the face of the impossibility of representing them digitally.

Book to book translation, the norm in paper-based technologies of editing, requires economies of scale that need not pertain in the digital; at the same time, the print edition's assumption of the materiality of the text it records dissolves within its own manufacture the bookishness of earlier forms. It is unsurprising, then, that print-determined editorial theories give prominence to ideas of texts as words rather than bodies; nor that New Bibliography, dominant in Anglo-American circles until the end of the 20th century, convincingly argued the paradoxical logic of eclectic text-editing as freeing textual forms from the inevitable corruption or failure of all their previous manifestations or bodies. By an equivalent paradox, the weightless and apparently limitless environment of the digital currently promises to anchor its e-texts in material objects, real rather than ideal textual bodies. The digital age is the age of the perceived significance of the material copy.

This, too, is less surprising than it might seem. What is facsimile editing but an investment in the reproducible social text? Both the digital and the book-historical turn can be traced to roughly the same moment around 1960 – Febvre and Martin's *L'apparition du livre* (1958) and Marshal McLuhan's celebration of the cool participatory media of electric communication (*The Gutenberg Galaxy*, 1962) set to liberate society from the conformities of print. The birth and death of print, then, circa 1960. McLuhan, McKenzie, and McGann continued to privilege the eventfulness and expressiveness of communication as conceived in the improvisational and participatory countercultures of the 1960s and 1970s – an understanding of print and electric textual forms alike as hippy 'happenings'. Witness McKenzie's McLuhan-haunted phrase 'forms effect meaning', while in McGann's late convergent thinking, only digital tools offer the means to 'develop a model for editing books and material objects rather than just the linguistic phenomena we call texts' (McGann 2013, 281).

The recovery and presentation of text as object, which I see as symptomatic of a particular digital editorial mindset, has obvious associations with curation. Of course, in its function to preserve and transmit cultural heritage, the act of curation always was implied in the work of editing. In its present digital form, however, text curation shares some of the implications of curation in its popular reconception as curationsim, encounter, or installation – not so much the exhibition of a discrete

object, as its situated reproduction as performance or interactive experience, expertly managed in the user's favour – a new participatory art that, thanks to visualization tools, pushes against previous limits on the acts of looking and consumption. Digitization releases both the potential for image customization and its userly affectiveness.

The new encounter challenges the traditional authority of the editor and raises the editorial stakes. As long ago as 1981, Randall McLeod wrote of the 'un-editing' potential of paper-based photofacsimiles of the earliest print editions of Shakespeare: 'Our editorial tradition has normalized text; facsimiles function rather to abnormalize readers' (McLeod 1981, 37). I take this to mean that facsimiles provide the means to make readers more critical by alerting them to the insecurity of certain conventions. Facsimile evidence challenges the editorial model in fundamental ways. Not least, now that the document underlying the new edition has been raised in status, its presence within the edition argues against or at least weakens the case for certain kinds of editorial licence. The document or object becomes more than a vehicle for text, it appears to become meaningfully indivisible from it.

Facsimile editions are undoubtedly a major benefit of digital editing. In revising our understanding of text and object they issue new critical challenges. We accordingly should be wary of suggesting that the 'un-editing' initiated by the facsimile does not require an even closer (and more suspicious) critical engagement with the textual object. By extension, something interesting is going on in our present refiguring of the print/non-print distinction that we might trace back to the sociological (or was it theological) textual turn established by McLuhan-McKenzie-McGann. On the one hand, we now see the book as no longer an inert object, a mere vehicle for text, but like the text it carries amenable to interpretation as a set of human actions, intentions, and interactions which are precious and specific; indeed, precious because they are specific, time-stamped and non-transferable. Meaning, we are happy to say, is not a privilege of text (which can be transmitted) but a property of things (which cannot), of objects. Under pressure from a sociologically inflected model of literary production we have re-ordered the forensic tools of our editorial trade. At the same time, we have become curators of objects and find their sites of particularity replicable within the electronic medium, and we are beginning to articulate arguments why this is so – why book objects might (even must) be digitally replicated – why the copy of the copy is the original.

References

- Greg, W. W. 1966. 'What is Bibliography?' In *Collected Papers*, 75-88. Oxford: Clarendon Press.
- McGann, Jerome. 2013. 'Why Digital Textual Scholarship Matters.' In *The Cambridge Companion to Textual Scholarship*, edited by Neil Fraistat and Julia Flanders, 274-288. Cambridge: Cambridge University Press.
- McKenzie, D. F. 'What's Past is Prologue.' In *Making Meaning*, edited by Peter McDonald and Michael Suarez, 259-75. Amherst: University of Massachusetts Press.
- McLeod, Randall. 1981. 'Un 'Editing' Shakespeare.' *SubStance* 10: 26-55.
- Obrist, Hans Ulrich. 2011. *Everything You Always Wanted to Know about Curating*. Berlin and New York: Sternberg Press.

The Videotext project

Solutions for the new age of digital genetic reading

Georgy Vekshin & Ekaterina Khomyakova¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The subject of this paper is the main communicative background and first results of the Videotext digital editorial project (<http://videotekst.ru/>), aimed to design the digital environment for tracing the author's text creation process in a reading mode.

Reading as activity. Hypertext... What is new?

It is well known, that reading as an activity (one of the organizing principles of modern social and cultural being) has changed considerably in the digital age. The linear (successive, syntactic) reading, upon which the idea of codex technology was based, has given way to a 'jumping' mode of reading, switching the eye from one page to another in a sufficiently random order. What was marginal for traditional books – its vocabulary, reference, hypertextual component – has become the biggest advantage of e-books. However, hypertext is not a product of the digital age, which only has generalized this principle. Hypertext, for all its importance for the reading and publishing culture, only rebuilds the reading process and makes it more complicated, but it does not present a new way of reading.

¹ philologos@yandex.ru

Textual criticism and the prospects of digital genetic reading. Linearity regained?

The storing and editorial representation of the main materials of textual criticism (manuscripts, pre-textual witnesses, and textual Versions – either as individual documents or aggregated and ordered) is only one possible approach to reading these materials. For the traditional book a natural way of organizing genetic materials always has been the hypertext. In this sense, textual criticism could be considered merely a mother of hypertext, and it would be strange if it does not use its capabilities in the digital age. However, it would be also strange if pre-textual information in digital editions, including scholarly editions, has limited the capabilities of its submission by a hypertext form, which is mostly discrete and does not provide a holistic perception of avant-text as reading object. At the moment, the main task may be the development of the technology for such an editorial digital representation of avant-text, that would be adequate to its nature, and that would provide the linearity of avant-text as a continual trail of the creative process.

As the new round of publishing and reading practices the syntactically arranged avant-text, showing the changes by the *measures of text animation – a digital technology for genetic reading* – can return to the process of reading its mostly lost linearity, and allow the reader to consciously follow the changes in any process of writing – to observe the formation of the text in its continuity. And together with the observed nascent text, of course, we must be able to follow the writing person – alive, acting and exhibiting himself and organizing himself in the creative workflow. Thus, the avant-text in its genetic dynamic representation should become the object of linear reading. Meanwhile it is mostly the heritage of textual critics, philologists, and only the most meticulous readers. New digital technologies can make avant-text a complete reading object.

The Videotext Project: goals, backgrounds and challenges

The most significant projects for digital modeling of the writing process can be considered as follows: *Juxta Software* (<http://www.juxtasoftware.org>), developed by NINES scholars; *Codex Sinaiticus* (<http://www.codexsinaiticus.org/en>); the ‘Handwritten Monuments of Ancient Rus’ (<http://www.lrc-lib.ru>); the online edition of V. Van Gogh letters (<http://vangoghletters.org/vg>), The David Livingstone Spectral Imaging Project (<http://livingstone.library.ucla.edu>), The Beckett Digital Manuscript Project (<http://www.beckettarchive.org>), The Proust’s notebooks digital edition (http://research.cch.kcl.ac.uk/proust_prototype/) and some others.

The computer-based tracing of the actual text production process, useful for literary genetic studies, can be carried out by *Scriptlog* (University of Lund, www.scriptlog.net), *Inputlog* (University of Antwerp, www.inputlog.net), *Translog* (Copenhagen Business School, www.translog.dk) and some other software.

The first attempt to animate drafts by cinematic means was made in Soviet Russia, where the outstanding textologist Sergey Bondi and screenwriter Sergey Vladimirsky removed a popular scientific film *The Pushkin’s manuscripts* (Mostekhfilm studio,

1937, rebuilt in 1961 <https://www.youtube.com/watch?v=FzHAdvGPj0E&t=6s>). The audience of the film was invited to follow the poet's work on the introduction to 'The Bronze Horseman' poem, tracing the appearing actual manuscript lines, crossings, replacements etc., backed by the voice of speaker, who announced the transcription. However, the cinematic animation is an energy-intensive laborious work, and it will never give the possibility to present a large amount of material, and the more to process it automatically. It is not interactive and cannot be used for further organizing, storing, studying, user's commenting and transforming the pre-text information. The task of that kind also could not be achieved by the animating 'kinetic typography' tools.

What concerns the attempts to digitally connect text versions in their dynamic sequence or to create animation tools to compile them within a readable 'live' presentation of text genesis – we know nothing about such an experience.

The 'Videotext' project, developed by Moscow researches and designers (G. Vekshin, leader of research group and the author of the model, E. Khomyakova, M. Gertzev and others) aims to create an online editorial system for kinetically visualizing genetic pathways of a literary and any other text. Initiated in 2011, it is still a work in progress, but already shows its potential as a valuable aid for scholars, tutors, editors and all software and internet users in their following the textual changes and making comments to avant-text from manuscript to printed copies. The XML format of Videotext software is compatible with most common browsers and offers a set of features for animating textual graphics to display dynamic genetic reading mode.

Alexander Pushkin's 'Arzrum Notebook' poems currently have been presented with Videotext software and are published as an experimental scholarly digital edition (<http://arzrum.videotekst.ru/>), including videotexts with textual comments, main editorial variants, manuscripts, transcripts and lifetime publications of poems.

We are working now on a specially designed semiotic system of moving graphic symbols and effects, named Video Typography, which allows to reflect basic text operations (appearance, disappearance, displacement, replacement) – with different types of animation; and further on to symbolize the details of the writing process (conjugated substitutions, variants not included in the manuscript, etc.) and 3 basic reasons for text changes, which can be emphasized with static markers (color, font style and so on).

The Videotext group also is working now on the compatibility of the software with automatic syntax, phonics, meter analyzers and so on, which will allow not only to make animated avant-texts, but also to provide visualizing and further automatic processing of huge text material.

Text, which is not *ergon*, but *energeia*, is always the embodiment of personality in its dynamics. We hope that by dynamically presenting the development of a text instead of merely juxta positioning its variants, a new range of possibilities for genetic reading is opened up. Instead of unquestioningly consuming, by 'reading from within' any user will be able to experience a text as source and result of a living process and co-create the text together with its original author.

A stemmatological approach in editing the Greek New Testament

The Coherence-Based Genealogical Method

*Klaus Wachtel*¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Paul Maas is one of the first scholars who used the term 'Stemmatik', English 'stemmatics' or 'stemmatology', albeit in inverted commas. In a 1937 article, 'Leitfehler und stemmatische Typen' he apologises for using this neologism and defines it en passant as 'Lehre von den Abhängigkeitsverhältnissen der Handschriften', 'Doctrine of dependencies between manuscripts'. The concluding statement of the article sets a high standard for an acceptable *stemma codicum*:

*As in a chemical formula the order of atoms is unambiguously and invariably determined for each molecule, so in a stemma the dependencies of witnesses for each passage of the text – if we are dealing with a virginal transmission. Against contamination there still is no remedy.*²

Two features have to be emphasized for a correct understanding of this doctrine:

1. It is restricted to 'jungfräuliche Überlieferung', virginal transmission.
2. It is about the relationship of manuscripts, not of the texts they carry.

1 wachtel@uni-muenster.de.

2 "Wie in der chemischen Formel die Anordnung der Atome für jedes Molekül einer Verbindung eindeutig und unveränderlich festgelegt ist, so im Stemma das Abhängigkeitsverhältnis der Zeugen für jede Stelle des Textes – wenn jungfräuliche Überlieferung vorliegt. Gegen die Kontamination ist noch kein Kraut gewachsen" (Maas 1937, 294).

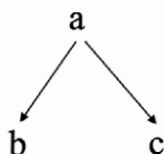
Maas first published his *Textkritik* in 1927. Giorgio Pasquali wrote a lengthy review about this work (1929) from which his *Storia della tradizione e critica del testo* (1934) developed. The overwhelmingly well documented conclusion of the review was that there is no immaculate manuscript tradition. Contamination is the rule, not an exception in rich manuscript traditions.

On the second point, the subject matter of stemmatology, I can cite Barbara Bordalejo from her recent programmatic article ‘The Genealogy of Texts: Manuscript Traditions and Textual Traditions’ (2016). The title already points to the importance of distinguishing manuscript and text, and Barbara duly emphasises the importance of this distinction for assessing traditional and computer-assisted stemmatology. The aim of conventional stemmatology is the reconstruction of the manuscript tradition. Gaps, i.e. losses, are replaced by hyparchetypies, and ideally the first manuscript of the transmission, the archetype, can be reconstructed. Here, however, conventional stemmatology also turns to the reconstruction of a text which then is printed as the text of an edition. The purpose of the whole enterprise would be pointless, if the archetype was preserved. Computer-assisted stemmatology focuses on the text carried by the manuscripts from the start. This is also true for the Coherence-Based Genealogical Method (CBGM) which was developed by Gerd Mink at the Institut für Neutestamentliche Textforschung in Münster. To use the words of Gerd Mink, ‘The CBGM deals with texts, not with manuscripts. The text is the witness’ (Mink 2009, 38).

The transmission of the Greek New Testament has several features which show that conventional stemmatology could not cope with it and in fact never tried. There is an overwhelming mass of medieval manuscripts which from the 9th century for the most part contain a standardised Byzantine text. This text, although clearly distinguished from older text forms in many places, developed different varieties, and these varieties often contain or, to use the pejorative term, are contaminated with variants from older text forms. Then there are quite a few witnesses from centuries before the 9th, the most comprehensive of which are from the 4th/5th centuries. Our earliest manuscripts, more or less fragmentary papyri, are dated to the 2nd-4th centuries. The early witnesses often share variants by which they are distinguished from the Byzantine text, but compared with each other they show many more differences than the Byzantine witnesses. This is due to the fact that most manuscripts from the first millenium are lost. What we have are single survivors, which support different text forms in constantly varying combinations. This means the NT tradition is highly, some say hopelessly, contaminated.

However, the Coherence-Based Genealogical Method, developed in the context of the *Editio Critica Maior* of the Greek New Testament, does offer a remedy against contamination. The remedy is the result of an analysis and interpretation of *coherence* which is the feature of the NT tradition balancing contamination or mixture.

Starting from an assessment of the genealogy of variants assembled in a full critical apparatus coherent structures are traced in the manuscript tradition. The assessments of variant passages are expressed by *local stemmata of variants*. A simple example may look like this:



If we make a statement about the relationship between variants as preserved in manuscript texts, we make a statement about the relationship between the texts containing these variants at the same time. If we say that variant a is prior to b and c, we also say that this passage is an instance of variation where the states of text containing a are prior to those containing b and c.

Hence the following principle of the CBGM:

A hypothesis about genealogical relationships between the states of a text as preserved in the manuscripts has to rest upon the genealogical relationships between the variants they exhibit. Therefore a systematic assessment of the genealogy of these variants (displayed as local stemmata) is a necessary requirement for examining the genealogy of textual witnesses (Gäbel et al. 2015, 1).³

Having constructed local stemmata for each variant passage we will be able to say in how many instances witness X has the prior variant as compared with witness Y at the places where they differ. As we are dealing with a contaminated tradition, there also will be a number of instances where Y has the prior variant. Finally there will be a number of unclear cases where the variants of X and Y are not related directly or it has to be left open which is the prior one. These numbers are tabulated in tables of potential ancestors and descendants.

Potential Ancestors of witness 35 (W1)

For the construction of a global stemma of witnesses a methodological step is necessary which defines an optimal substemma for each witness. An optimal substemma comprises only the ancestors that are necessary to account for the individual text of a witness. Gerd Mink's online 'Introductory Presentation' to the CBGM contains a very clear explanation of the procedure leading to an optimal substemma (2009, 164-177).

In the Introductory Presentation Mink demonstrates the procedure in detail for 35. As a result, four witnesses emerge, 617, 468, 025 and 1739. A global stemma

3 The basis for this comprehensive analysis is a critical apparatus comprising all variants of every Greek textual witness selected for the edition.

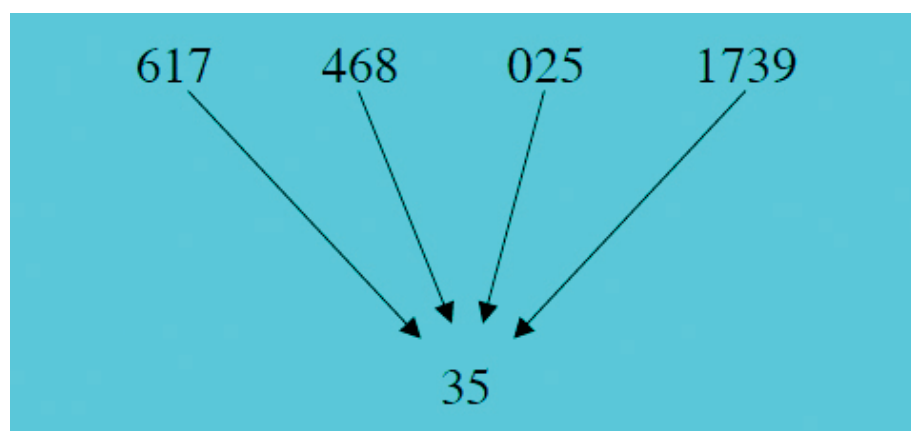
W2	NR	D	PERC1	EQ	PASS	W1<W2	W1>W2	UNCL	NOREL
617	1		95.995	2924	3046	56	46	15	5
424	2		95.988	2919	3041	51	45	23	3
468	3		95.588	2903	3037	57	49	27	1
A	4		92.263	2695	2921	212	0	6	8
025	5		91.160	2444	2681	99	84	46	8
323	0	-	89.638	2725	3040	111	111	76	17
1739	6		87.853	2676	3046	158	115	77	20
03	7		87.272	2633	3017	201	78	90	15
04	8		87.262	1836	2104	103	93	60	12
P74	0	>	82.493	278	337	27	22	6	4

Version 1.0

W2: Manuscript numbers of potential ancestors -- NR: Ranking numbers according to degrees of agreement -- PERC1: Percentage of agreement with 35 (=W1) -- EQ: Number of agreements with 35 -- PASS: Passages shared by 35 and W2 -- W1<W2: Number of priority variants in W2 -- W1>W2: Number of priority variants in 35 (=W1) -- UNCL: Unclear relationship between W1 and W2 -- NOREL: No relationship between W1 and W2.
(Data Source: Cath. Letters (excl. small fragments and extracts)).

consisting of optimal substemmata will meet the condition that Maas formulated for a stemma based on an immaculate manuscript tradition. It has to be true at each variant passage. This means that the stemmatic ancestors of each witness will be either equal or support a variant prior to the one in the descendant.

Optimal Substemma for 35



A Remedy against contamination

1. Do not try to reconstruct the manuscript tradition but focus on the states of text preserved in the manuscripts.
2. Do not try to use joining and disjoining errors (*Binde- und Trennfehler*) to define strands of transmission, because *diorthosis* is an integrated part in the manual reproduction of texts. Instead, base your research on a well constructed apparatus of grammatically valid variants and use tabulated rates of agreement (pregenealogical coherence) to determine the proximity of each state of text.
3. Do not try to reconstruct hyparchetypes, because for all richly transmitted texts from antiquity you would need several generations of them, piling up sub-hypotheses about stages in the development of the text. Instead, trace the structures in the preserved states of text and order them according to rates of agreement (pregenealogical coherence) and according to the genealogy of the variants contained in them (genealogical coherence).
4. To explain the development of the text as documented in extant witnesses (i.e. states of text) determine for each of them their *stemmatic* ancestors, i.e. the ancestors needed to explain the respective state of text by agreement and descent. Read bottom up, the resulting stemma will account for the reconstruction of the initial text at its top.

References

- Bordalejo, Barbara. 2016. 'The Genealogy of Texts: Manuscript Traditions and Textual Traditions.' *Digital Scholarship in the Humanities* 31: 563-577.
- Gäbel, Georg, Annette Hüffmeier, Gerd Mink, Holger Strutwolf and Klaus Wachtel. 2015. 'The CBGM Applied to Variants from Acts – Methodological Background.' *TC Journal* 20. <http://rosetta.reltech.org/TC/v20/TC-2015-CBGM-background.pdf>.
- Maas, Paul. 1937. 'Leitfehler and stemmatische Typen.' *Byzantinische Zeitschrift* 37: 289-294.
- . 1957. *Textkritik*. Teubner: Leipzig. 3rd edition.
- Mink, Gerd. 2009. 'The Coherence-Based Genealogical Method (CBGM) – Introductory Presentation.' <http://egora.uni-muenster.de/intf/service/downloads.shtml>.
- Barbara, Kurt Aland†, Gerd Mink, Holger Strutwolf, and Klaus Wachtel (eds). 2013. *Novum Testamentum Graecum: Editio Critica Maior*. Deutsche Bibelgesellschaft: Stuttgart. 2nd edition.
- Pasquali, Giorgio. 1929. 'Rez. Maas: Textkritik.' *Gnomon* 5: 417-435, 498-521.
- . 1971. *Storia della tradizione e critica del testo*. Firenze. 2nd edition.

WP2

Technology, Standards, Software

What we talk about when we talk about collation

Tara L. Andrews¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

In 2010 a roundtable on digital methods for philology was hosted in Leuven, at which I was asked to describe what a digital workflow for critical edition of texts might look like.² During the discussion that followed, it became clear that different scholars have very different views of what constitutes 'a collation'. Although the acts of transcription and collation often are regarded as separate steps in digital workflows for critical editions, many textual scholars regard the collation as a distinct entity in its own right, comprising the text of the individual witnesses and the correspondence between them, inseparable from the acts that go into its creation.

The purpose of this contribution is to discuss the various definitions of 'collation' that have arisen in textual scholarship. This is particularly relevant at a meeting of the ESTS devoted to digital methods of textual scholarship, in an environment that has witnessed the development of software to carry out what is known as 'automatic collation'. If tool developers and digitally-minded textual scholars have an understanding of what collation is that differs from that of less digitally-minded peers, this difference needs to be drawn out so that fruitful collaboration may continue. This contribution will also touch on the use of collation software and other such 'black boxes' and their relation to what we think of as scholarly pursuit.

1 tara.andrews@univie.ac.at

2 The paper that arose from the round table talk has since been published (Andrews 2012).

Defining collation

Handily enough, the *Lexicon of Scholarly Editing* has brought together several definitions of collation, culled from various sources. Examining these in more or less chronological order, we find that the concept of what a collation is has evolved, and varied, according to the aims of the editor whose definition is used and according to the capabilities of the time. Into and beyond the 1960s, one conceived of a collation as a process carried out with reference to a base text, usually some kind of norm such as a published edition (Colwell and Tunc 1964, 253). By the early 1990s, perhaps spurred on by the adoption of computer technology and the relative ease of whitespace tokenization and pairwise comparison, collation was described as the comparison of ‘two genetic states or two versions...of a text’ (Grésillon 1994, 242) and something that was done ‘word for word’ (Stussi 1994, 123), albeit still with respect to a reference text. Computational affordances could be carried yet farther, with a further definition of collation as an act that was carried out ‘character for character’ (Shillingsburg 1996, 134). This definition is striking in another aspect: rather than referring to comparison with a base text, its author calls for the comparison of ‘all versions that could conceivably have been *authoritatively* revised or corrected.’ It is around this time that the notion of the base text ceases to be a central part of the definition of the collation. Later scholars define collation as an act whose purpose is to find agreements and diversions between witnesses (Plachta 1997, 137) or explicitly to track the descent of a text (Kline 1998, 270); differentiate between collation as a process of comparison (carried out ‘word-for-word and comma-for-comma’) and the result of comparison that is known as the ‘historical collation’ (Eggert 2013, 103); or describe collation again as a process, the result of which is described simply as lists of variant readings (Greetham 2013, 21).

From these descriptions, it is possible to detect a converging (though also evolving) definition of collation, and a distinction between the act and its result. Collation may be carried out against a reference text, pairwise, or as a many-to-many comparison. The comparison may be done at the word level, at the character level, or at another unspecified syntactic or semantic level, according to the sensibilities of the editor. The question of authority enters into a collation: the witnesses to be compared should have some substantive claim to influence the editor’s idea of what the text is. The purpose of collation is usually given as being the discovery of where witnesses to a text converge and diverge; one might also claim that its purpose is to track the descent or the genesis of a text.

The act of collation produces a result, also known as a collation. Greetham (1994, 4) refers to the *apparatus criticus* and historical collation as a representation of a ‘collation and the results of emendation’. From this we may deduce the definition of collation-as-result, the thing often known as the ‘historical collation’: this is the *apparatus criticus* of an edition minus any emendations. It is important to note here that this historical collation is almost always a curated and pruned version of the results of comparison of the text, a fact to which Eggert (2013, 103) also alludes when he writes that the historical collation “is often restricted to... ‘substantives’, leaving the now-orphaned commas and other ‘accidentals’ to look after themselves.” In that sense the collation, as many textual scholars understand

it, is a document that reflects not only the ‘raw’ results of comparing a text, but also the scholarly work of interpreting these results into a particular argument about the constitution and history of that text.

Automatic collation

The author of one of the first well-known text collation tools was initially taken with ‘the notion of feeding these manuscripts into one end of the computer, which would then extrude a critical apparatus on the other’ (Robinson 1989, 99). His tool, COLLATE, was designed to work interactively and closely with the editor, not only to support the process of collation but also to produce the scholarly artefact that we call a collation.

The current generation of collation tools, on the other hand, limit themselves strictly to the process; the authors of the CollateX tool describe collation simply as text comparison and refer to it as a process (Dekker *et al.* 2014, 453). The process of collation around which these tools are based, also known as the ‘collation workflow’, is known as the ‘Gothenburg model’ after its definition there at a workshop in 2009. These discrete steps of tokenization, normalization, alignment, and analysis form the process by which a scholarly collation artefact is generally produced.

There are three pieces of software currently widely available known as ‘collation tools’; according to the Gothenburg model each of these is a tool for alignment, although some include other capabilities. Each has a different model for alignment, conforming to the models we have seen in our definitions of collation above. TuSTEP compares witnesses against a selected base text; JuXta compares pairs of texts, and can visualize the aggregate results obtained by pairing each witness against a selected reference text; CollateX performs a many-to-many comparison of all witnesses without a base text. In none of these tools does the result of this comparison amount to what a scholar would recognize as a ‘historical collation’, or an ‘*apparatus criticus* without the emendations’. Rather, it remains the task of the editor, whether supported by additional software tools or not, to perform the philological analysis on these comparison results in order to produce the ultimate collation and, thereby, the edition.

Collation and the black box

Textual scholars often treat the idea of automatic collation with considerable suspicion. Those who do prepare their editions using these tools often encounter a mix of fascination and resistance among their colleagues, who on the one hand understand the utility of the digital medium, but on the other hand cannot bring themselves to trust software to produce a collation. We now can see how the definitional divergence described here might contribute to this lack of trust. If a textual scholar regards a collation as an artefact that reflects not only straightforward comparison of textual witnesses but also a selection and adaptation of those comparison results that is used to put forward an argument about the genesis or tradition of that text, it is both reasonable to wonder how a piece of

software might do all that, and also to worry that the software is being given an authority over scholarly interpretation that the editor is being told not to question.

Automatic collation, however, is a different sort of black box. The tool compares, whether its user understands its methods of comparison or not, and it produces a result. The authority delegated to the tool is not to judge whether a given alignment is 'right' or 'wrong', but rather, and only, to align readings. Modification, interpretation, and even dismissal of that result remains the prerogative of the editor who obtains it.

References

- Andrews, Tara L. 2012. 'The Third Way: Philology and Critical Edition in the Digital Age.' *Variants* 10: 1-16.
- Colwell, E. C. and E. W. Tune. 1964. 'Variant readings: classification and use.' *Journal of Biblical Literature*, 83: 253-261.
- Dekker, Ronald H., Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert. 2015. 'Computer-supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project.' *Digital Scholarship in the Humanities* 30. 3: 452-470.
- Eggert, Paul. 2013. 'Apparatus, text, interface.' In *The Cambridge Companion to Textual Scholarship*, edited by Neil Fraistat and Julia Flanders, 97-118. Cambridge: Cambridge University Press.
- Greetham, David C. 1994. *Textual Scholarship. An introduction*. New York & London: Garland Publishing.
- . 2013. 'A History of Textual Scholarship.' In *The Cambridge Companion to Textual Scholarship*, edited by Neil Fraistat and Julia Flanders, 16-41. Cambridge: Cambridge University Press.
- Grésillon, Almuth. 1994. *Eléments de critique génétique lire les manuscrits modernes*. Presses universitaires de France.
- Kline, Mary-Jo. 1998. *A Guide to Documentary Editing. Second Edition*. Johns Hopkins University Press.
- Plachta, Bodo. 1997. *Editionswissenschaft: eine Einführung in Methode und Praxis der Edition neuerer Texte*. Stuttgart: P. Reclam.
- Robinson, Peter. 1989. 'The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation.' *Literary and Linguistic Computing* 4: 99-105.
- Shillingsburg, Peter L. 1986. *Scholarly Editing in the Computer Age: Theory and Practice*. The University of Georgia Press.
- Stussi, Alfredo. 1994. *Introduzione agli studi di filologia italiana*. Bologna: il Mulino.

The growing pains of an Indic epigraphic corpus

Dániel Balogh¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

This paper introduces a recent initiative in digital epigraphy under the aegis of the ERC Synergy project 'Beyond Boundaries – Religion, Region, Language and the State.'² The project as a whole aims to re-evaluate the social and cultural history of the Gupta period (loosely defined as the 4th to 7th centuries CE) in South, Central and Southeast Asia. Studies of this region and time largely have been compartmentalised, for instance on the basis of disciplines and of modern-day political states. In order to transcend these boundaries and approach an understanding of the region as an interconnected cultural network, 'Beyond Boundaries' involves scholars of several disciplines (such as archaeology, cultural history, numismatics and philology) and various regional/linguistic foci (e.g. India, Tibet and Burma), affiliated to one of the host institutions – the British Museum, the British Library and the School of Oriental and African Studies – or contributing from other institutions across Europe.

One component of this project is the 'Siddham' database of Indic epigraphic texts.³ Its development was commenced in the summer of 2015 with the encoding of previously published Sanskrit inscriptions created under the imperial Gupta rulers. It will be expanded progressively both horizontally (by adding inscriptions

¹ danbalogh@gmail.com.

² www.siddham.uk. Project number 609823; <https://asiabeyondboundaries.org/about/>

³ www.siddham.uk. *Siddham* (Sanskrit: 'accomplished, fulfilled, perfected') is the first word of many Indic inscriptions. It basically serves as an auspicious invocation, but may in at least some epigraphs carry the more concrete meaning that an action the inscription commemorates (such as a donation of land or the construction of a building) has been carried out. The word 'Indic' is used here in a loose sense. Presently the corpus is comprised of epigraphs from the Indian subcontinent, mostly Sanskrit with some Prakrit (Middle Indic). It may, in time, be extended to non-Indo-Aryan languages such as Tamil, Tibetan and Pyu, and outside the subcontinent to Sri Lanka, Tibet and Southeast Asia.

from other dynasties and regions) and vertically (by accumulating metadata, gradually increasing the granularity of markup, and through re-editing crucial inscriptions). The public launch of the website is planned for the autumn of 2017. Our objective is to create a digital corpus that will be freely accessible and useful not only for epigraphists and specialists of the relevant languages but, by including translations of the texts, for scholars of other disciplines as well as for interested lay audiences. We are also aiming in the long run to develop an interface through which new inscriptions can be added by scholars anywhere with web access and without a need for lengthy training in our methods.

According to one of the 20th century's foremost experts on Indian epigraphy, as much as 80% of what we know about Indian history in the 1st millennium CE and before is derived from inscriptional sources (Sircar 1977, 91). Figures notwithstanding (after all, how does one weigh for instance the date of a king's accession against the ground plan of a temple founded by him?), epigraphic sources are outstandingly important for the study of the Indian subcontinent, given pre-modern India's proverbially casual approach to factual history. The fact that texts written on perishable traditional media such as palm leaf rarely survive for more than a few centuries in the Indian climate further underscores the significance of records in stone and copper.

Indic epigraphic studies have been pursued for over two hundred years, with a peak in the late 19th and early 20th century. The total number of known Indian inscriptions has been estimated variously from about 90,000 (Sircar 1977, 91) to 200,000 (Havanur 1987, 50), of which about 58,000 have been edited in accessible publications (Garbini 1993, 64). The overwhelming majority of this wealth is, as can be expected, relatively recent; the number of epigraphs increases almost exponentially as we approach modern times. For the timeframe of *Beyond Boundaries*, the tally is well below one thousand inscriptions, most of which range from a few dozen to a few hundred words.⁴

Although this scope is dwarfed in comparison to corpora of classical antiquity such as EAGLE, Siddham is still the most ambitious project to date in the field of digital Indic epigraphy. It is by no means the first, though, and 'the need for comprehensive computer databases of the now unmanageably vast published epigraphic material' has long been recognised as '(m)ost urgent' (Salomon 1998, 224). Earlier endeavours include digital versions of massive amounts of Sanskrit and Prakrit inscriptions and/or their translations, but these are presented with little to no structure and/or face serious accessibility problems as they become increasingly dated. A remarkable pioneer is the 'Indian Epigraphy' website⁵ created in the early 2000's by Dmitriy N. Lielukhine at the Oriental Institute of Moscow, which included a large number of texts carefully digitised (as opposed to simply dumping OCR output on the web) and presented in a structured way, but employed a custom character encoding that is only readable in tandem with

4 Short epigraphs consisting of just a few words are probably underrepresented in the published material on account of their meagre historical 'value' as perceived by the scholars editing the more substantial inscriptions.

5 Sadly defunct since 2014, the URL was <http://indepigr.narod.ru>. Much of the content is still accessible (with difficulty) via archive.org.

a font designed for the purpose. Later endeavours employing better technological solutions also exist, but are more limited in temporal and/or geographical scope.⁶

At present, the Unicode standard lets us put aside worries about compatibility for a long time to come. The British Library intends to provide for the continuing existence of the Siddham website beyond the lifespan of the mother project, although the issues of maintenance and of the acceptance and curation of future contributions are yet to be settled. To ensure that the essence of the corpus remains available regardless of the long-term survival of the website, the content will also be shared on GitHub in the form of XML files. The increasing currency of TEI in general and EpiDoc in particular will facilitate the interoperability and longevity of our creation.

EpiDoc⁷ serves as the flesh and blood of our corpus, as texts are stored in XML snippets, each comprising the `<div type='edition'>` of a full EpiDoc file. We use a hybrid input scheme that utilises bracket combinations for common features of the texts and only requires XML tags for structure and for rare textual features. This way the texts remain largely human-readable in their raw form, but can be converted automatically to full EpiDoc. This input method may also be used later on to allow scholars outside the project to contribute to the database without XML expertise. In addition to EpiDoc, the Siddham corpus has a skeleton consisting of relational database tables. The edition snippets, along with other snippets containing translations (and, optionally, critical apparatus and commentaries), are referenced from an 'Inscriptions Table' that additionally stores metadata pertinent to each inscription, such as layout and hand description, language and date.

A separate 'Objects Table' serves as the repository of metadata pertaining to inscription-bearing objects, such as physical properties (material, dimensions and freeform description) and history. Entities in the object table may have 'daughters' when an object consists of several physical parts (such as a set of copper plates, or a fragmented stone tablet). In such cases it is always an abstract 'whole' object entity that is linked to one or more inscription entities; the daughter objects may have individual physical descriptions and histories, but do not have links to inscriptions. (The boundaries of the component objects, however, may be shown in the edition of the text using milestone elements.)

The separation of object metadata from inscription metadata is conceptually desirable as it brings objects to the fore as entities in their own right rather than mere dismissible substrates of the texts they carry. It is also helpful in situations where a single object is home to multiple inscriptions. Each entity in the inscriptions table is paired to only one entity in the objects table, but object entities may be in a one-to-many relationship with inscription entities. Thus object data need not be iterated for each inscription on that object, and if object data are updated, this only needs to be done once. When, however, a corpus entry is exported as a full EpiDoc file, its TEI header will incorporate data from the objects table and thus

6 Notable examples are 'Sātavāhana Inscriptions' (<http://162.243.35.29:8080/exist/apps/SAI/index.html>) and the 'Corpus of the Inscriptions of Campā' (<http://isaw.nyu.edu/publications/inscriptions/campa/>).

7 EpiDoc is an application of TEI for encoding epigraphic documents. See <http://www.stoa.org/epidoc/gl/latest/>

will involve redundancy in the case of multiple inscriptions on a single object.⁸ A difficulty inherent in this demarcation is the placement of images and bibliographic references in the scheme. These, at the present conceptual stage, belong solely in the inscriptions table, but this still generates some redundancy when a publication or an image concerning the whole of an object must be included with each of the inscriptions on that object.

At the front end, users will be allowed to search text and to browse objects and inscriptions by various criteria. Object records will be displayed with links to the inscription(s) on that object, and inscription texts may be displayed in several alternative forms including diplomatic, edited and raw XML (with the former two generated by a transformation of the latter). Texts are stored and displayed in Romanised transliteration (IAST). This allows the addition of editorial spacing between words even where the syllabic organisation of their original Brāhmī-type script (and of Devanāgarī, the modern Indian script often used in printing Sanskrit) would prevent this.⁹

In addition to rendering the texts more accessible to scholars less proficient in the language, Romanisation and editorial spacing also facilitate the tokenisation of words for searching and referencing. However, this involves an additional hurdle that is most prominent in Sanskrit, the chief epigraphic language of India in the period we are concerned with. In the phenomenon of *sam̐dhi* (euphonic alteration), word-final phonemes may be merged with the following initial phoneme into a single phoneme that is not necessarily identical to either of the original ones. This feature of the language is reflected in writing, which makes it impossible to neatly wrap individual words in tags unless one is willing to truncate words arbitrarily.¹⁰ At the present stage this problem is ignored in the Siddham corpus since tokenisation is not on the agenda. Nonetheless, partial tokenisation (such as the tagging of personal and geographical names) may well be a medium-term goal, and in the long term full tokenisation may also become desirable. The problem of fuzzy word boundaries may be handled by reluctantly accepting truncation and alteration or by applying standoff markup; either of these solutions will need to be accompanied by lemmatisation.

8 A systematic approach to handling n to n relationships between objects and inscriptions in EpiDoc has been suggested by Morlock and Santin 2014; their method is more complex than the two-pronged approach outlined here, but may be adopted by Siddham for use in EpiDoc export.

9 In the 'abugida' scripts of South and Southeast Asia by and large each glyph represents a group of one or more consonants followed by one vowel. There are also glyphs to represent vowels in hiatus or at the beginning of blocks, but word-final consonants are as a rule joined in a single glyph to the initial vowel or consonant-vowel cluster of the following word. Although final consonants can be represented in these writing systems when necessary, the special signs for these are hardly ever used except at the end of major structural or semantic units.

10 Thus printed editions of Indic epigraphic texts commonly show the merged phoneme as the beginning of the latter word, which leaves the former word invariably truncated and the latter word's initial phoneme frequently altered.

References

- Garbini Riccardo. 1993. 'Software Development in Epigraphy: Some Preliminary Remarks'. *Journal of the Epigraphical Society of India* 19: 63-79.
- Havanur, S. K. 1987. 'Analysis of Inscriptional Data Through Computer'. *Journal of the Epigraphical Society of India* 14: 50-55.
- Morlock, Emmanuelle, and Santin, Eleonora. 2014. 'The Inscription between text and object: The deconstruction of a multifaceted notion with a view of a flexible digital representation'. In *Information Technologies for Epigraphy and Cultural Heritage (Proceedings of the First EAGLE International Conference)* edited by Silvia Orlandi, Rafaella Santucci, Vittore Casarova, and Pietro Maria Liuzzo, 325-350. Roma: Sapienza Università Editrice.
- Salomon, Richard. 1998. *Indian Epigraphy*. New York / Oxford: Oxford University Press.
- Sircar, Dinesh Chandra. 1977. *Early Indian Numismatic and Epigraphical Studies*. Calcutta: Indian Museum.

The challenges of automated collation of manuscripts

Elli Bleeker,¹ Bram Buitendijk,²

Ronald Haentjens Dekker,³ Vincent Neyt⁴

& Dirk Van Hulle⁵

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

For some time now, the Huygens ING in Amsterdam and the Centre for Manuscript Genetics in Antwerp have been working together to find a good solution for the automatic collation of our transcriptions of modern manuscripts. The Huygens ING develops CollateX, and the transcriptions of Samuel Beckett's manuscripts made in Antwerp serve as a test case.

One of the problems we are facing is that Beckett often did not use different colours of ink when he made revisions, so it is difficult to discern writing layers with certainty, and with automatic collation in mind, it is impossible to divide a manuscript up into multiple coherent textual versions, one for every layer of revision. This problem is of course not unique to modern manuscripts; the phenomenon also occurs in medieval and older documents. And actually, we prefer the word 'challenge' to 'problem', as we see the complexity of this textual form as a richness that deserves to be treated with attention to detail, also in collation.

In the last few months, we have worked together on the development of a new version of CollateX that can handle this kind of internal variation or 'in-text variation', if it is encoded in XML using the common practices described in the TEI guidelines. Our goal is to lower the threshold for CollateX users to input their

1 elli.bleeker@uantwerpen.be.

2 bram.buitendijk@huygens.knaw.nl.

3 ronald.dekker@huygens.knaw.nl.

4 vincent.neyt@uantwerpen.be.

5 dirk.vanhulle@uantwerpen.be.

TEI/XML transcriptions, and to build special mechanisms into the algorithm to preserve the multi-layered richness of the manuscript (as much as possible) in the output of the automatic collation.

The TEI guidelines propose two ways of encoding in-text variation. On the one hand, added text can be encoded with the <add> tag, cancelled text with the tag. There is the option of marking a cancellation and its following addition as one substitution with the <subst> tag. On the other hand, in-text variation can be encoded using the apparatus (or <app>) tag. This approach means, to quote the TEI guidelines, ‘treating each state of the text as a distinct reading’ (TEI).

The latter method was used among others in the encoding system developed by Barbara Bordalejo for the *Commedia* project (Bordalejo 2010). She adds an extra dimension to this apparatus tagging, by making a clear distinction between ‘the text of the document’ and ‘how the editor (or the transcriber) interprets the different stages of development of the text’. The two ‘textual states’ produced by a substitution are made explicit as readings, and a third reading with the type attribute value ‘literal’ is used to transcribe ‘the visible, physical features of the text of the documents’ (Bordalejo 2010), as shown in an example below.

```
<app>
    <rdg type="orig">dura</rdg>
    <rdg type="cl">duro</rdg>
    <rdg type="lit">dur<hi rend="ud">a</hi>o</rdg>
</app>
```

In the two encoding schemes, the <app> tag and the <subst> tag perform the same function: they flag and demarcate a spot where a manuscript becomes ‘multi-layered’, where there are two or more alternatives for one word or passage. The author usually discards the cancelled alternative in the next version of the text, but to facilitate the examination of the writing process we believe it is important to include the cancellations in the collation input as well, and to have them clearly visualized as such in the output.

In order to make this possible in a collation tool such as CollateX, the first step must be to allow for the comparison of structured data, such as XML. Currently, XML tags can be passed along through the collation pipeline, but usually they do not influence the alignment of tokens. Our proposal is based on a proof of concept where a witness inputted as XML will be treated as a tree hierarchy containing text, elements, attributes and attribute values. All this information will be taken into account during the collation.

The new collation algorithm is designed to process the <subst> as well as the <app> tagging scheme. They both trigger special behaviour in the algorithm: CollateX labels <subst> and <app> as an ‘OR-operator’, and the <add>s, s and <rdg>s as ‘OR-operands’. This means that the elements and the text they contain are not to be placed in a linear sequence of tokens, but are seen as two options for

the same spot in the sequence. In this way they do not need to be aligned next to each other, but possibly above and below each other.

It is important to stress that the special treatment of transcriptions containing `<add>`s and ``s only works when `<subst>` tags have been placed explicitly around them. Loose `<add>` and `` tags will be treated as all other tags in the input transcription. In the case of the `<app>` method, an additional rule of exception has been created for the ‘literal reading’ `<rdg type=‘lit’>` used in the *Commedia* project. The new CollateX algorithm disregards the reading with the ‘lit’ attribute value and includes all other readings in an `<app>` in the collation. At the moment of writing, we are in the midst of development. We already have a working implementation of this OR-operand, but the software can only process 2 witnesses that have no other XML elements than the `<subst>` and `<app>` tagging schemes. This is, of course, temporary: in the end the software should be able to process multiple witnesses.

Standard substitution

Here is the XML input for a standard substitution using the `<subst>` tagging (1) and the corresponding input using the `<app>` tagging scheme (2):

1. Bench by the

```
<subst>
    <del hand="#SB">lock</del>
    <add hand="#SB">weir</add>
</subst>
```

2. Bench by the

```
<app>
    <rdg type="deletion">
        <del hand="#SB">lock</del></rdg>
    <rdg type="addition">
        <add hand="#SB">weir</add></rdg>
    <rdg type="lit">
        <hi rend="strike">lock</hi>
        <hi rend="sup">weir</hi></rdg>
</app>
```

Please note how the <rdg type='lit'> contains only 'documentary' information, and the first two readings contain the interpretative tagging, such as <add>, and all possible attributes associated with these tags. It is precisely this interpretative tagging we as (genetic) editors would like to pass through the collation process and include in the apparatus generated by CollateX.

When collated with a second witness 'Bench by the weir', CollateX will produce the following output (here expressed in the TEI parallel segmentation format), exactly the same for both input methods:

Bench by the

```
<app>
  <rdg wit="#Wit1" type="deletion" varSeq="0">
    <del hand="#SB">lock</del>
  </rdg>
  <rdgGrp type="tag_variation_only">
    <rdg wit="#Wit1" type="addition" varSeq="1">
      <add hand="#SB">weir</add></rdg>
    <rdg wit="#Wit2">weir</rdg>
  </rdgGrp>
</app>
```

When a witness contains a substitution, marked in the input with the <subst> tag or with the <app> tag, the siglum for that witness will be present two or more times in the readings of the <app> produced in the output. A @type attribute on the <rdg> element indicates whether it concerns a 'deletion', an 'addition', or an 'addition within a deletion', and so on. The varSeq or variant sequence attribute numbers these multiple Wit1 readings in the linear order in which they are placed in the input transcription. As most projects will place the deleted word before the added word in their transcriptions, the deletion will have varSeq equals zero, and the addition varSeq equals one.

When two witnesses have the same word (or words), but the use of the surrounding tags is different, they are placed in a reading group (<rdgGrp>). It has a type attribute with the value 'tag_variation_only' to indicate that the readings only differ on the level of the XML elements. In the example witness 1 and 2 have the word 'weir', but there is a difference: the occurrence of 'weir' in the first reading is contained in an <add> element, which is not the case in the second reading.

We acknowledge that this tagging is already very verbose for such a simple example. But it holds a lot of information, and that information can be transformed (for instance with XSLT) into any number of more human-readable visualisations, such as this alignment table:

Wit2	Bench by the	weir
Wit1	Bench by the	1 weir 0 lock

Witness one has in-text variation: at the ‘level zero’, i.e. in the running text, the word ‘lock’ has been struck through (the strikethrough being a visualisation of the tag in the output), and the word ‘weir’ has been added (with superscript as a visualisation convention for the <add>tag). The red colour draws attention to the matching reading of ‘weir’ in Witness 1 and Witness 2. We will provide an XSL stylesheet with a basic visualisation that then can be customized to every project’s desires.

In the last part of this extended abstract, a number of more complex textual instances will be discussed that often occur in manuscripts.

Substitutions within a word

Consider the example of the word ‘furthest’ being changed to ‘farthest’ by crossing out the ‘u’ and writing an ‘a’ above the line. There are quite a few ways of encoding something like this in both tagging schemes. But in the case of the <subst> method, the placement of the <subst> tag itself is problematic. Placing it around only the changed letters would produce an undesirable and unusable collation result. Placing it around the entire word makes more sense, but it complicates matters immensely at the tokenization stage in CollateX to handle these instances correctly. We currently are exploring ways to solve this issue. The problem does not arise with the encoding via the <app> tagging scheme, as the ‘literal’ reading can contain the letter for letter substitution, while the two interpretative readings can hold the two full words ‘furthest’ and ‘farthest’, optionally with and <add>tags around the modified characters.

Wit2	the	farthest
Wit1	the	1 f ^a rh u sthest 0 furthest

Alternative Readings

The same principle of the OR-operator applies to alternative readings or open variants. As they are not substitutions, the <subst> tagging scheme is not a suitable way of triggering the special treatment in CollateX. Our recommendation would be to use the <app> method here.

Transpositions

A transposition produces ‘two states of the text’, with a difference in word order. It is another instance of non-linearity in a manuscript, but it is of a different type. There are not two alternatives for the same ‘spot’ in the linear sequence, but two different linear sequences. A special treatment by CollateX of this textual feature is not strictly needed to provide a meaningful result. But editors can make use of it by fitting the transposition into the <app> tagging scheme.

Long substitutions

We would like to present our last example, a long substitution, to address a challenge we are facing in development.

Wit2	The dog's	big eyes	.
Wit1	The dog's	¹ brown eyes ⁰ big black ears	.

Wit1: The dog's big black ears brown eyes.

Wit2: The dog's big eyes.

Wit2	The dog's	big		eyes	.
Wit1	The dog's	big	¹ brown ⁰ black	¹ eyes ⁰ ears	.

In this case, there is more than one way to align all of the words in these witnesses. It has to do with the preferred focus in the collation result: do we want to give priority to the unit of the substitution, or should the alignment of matching words receive priority?

This alignment table draws full attention to the textual operation performed by the author in Witness 1: 'big black ears' was struck through and substituted by 'brown eyes'. It clearly marks the spot where 'something happens' on the manuscript and you get the totality of the textual operation as one block.

If we ask CollateX to give dominance to the matching of words, we arrive at this more detailed alignment table. The similarities jump out at you, but the unit of the substitution can get lost somewhat. An editor making an apparatus by hand is free to choose a different focus depending on the corpus or even alternate for different parts of the same corpus, but a software tool needs clear instructions, and we have not decided yet which instructions we would like to give.

In conclusion, in this paper we provide an inclusive approach to automated collation that can account for multiple types of textual variation. We are developing CollateX to recognize the two TEI XML tagging schemes for in-text variation and to label them as OR operators, thus treating multiple layers of revision in a more suitable way.

Our approach comprises a rather complex TEI/XML encoding of the collation input and output. However, what happens on the page of the source document is complex and the more we would simplify it by linearizing or flattening the input, the more information we lose. It is exactly this information we consider important for research into manuscripts and literary writing processes.

References

- Bordalejo, Barbara. 2010. 'VII. Appendices: C. The Commedia Project Encoding System'. <http://sd-editions.com/AnaServer?commedia+6215691+viewarticle.anv+printdoc=1>
- TEI Consortium. 'P5: Guidelines for Electronic Text Encoding and Interchange'. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html>

The role of digital scholarly editors in the design of components for cooperative philology

Federico Boschetti,¹ Riccardo Del Gratta²

& Angelo Mario Del Grosso³

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

This contribution is focused on the role of the digital scholarly editor in the continuous process of analysis, development and evaluation of libraries of components for cooperative philology. We need to distinguish between collaboration and cooperation, because these different practices have important implications on project design. Indeed, according to Kozar (2010), collaboration requires direct interaction (i.e. negotiations, discussions, etc.) among individuals to create a product, whereas cooperation requires that all participants do their assigned parts separately and provide their results to the others. As a consequence, collaborative philology produces web applications and platforms, so that digital humanists organized in communities can work together with shared goals (McGann 2005). On the other hand, cooperative philology produces web services and libraries of components that highly decouple their function from the overall goal they are used for. In cooperative philology, the interactions are not limited to person-to-person transactions, but involve the cooperation between human and non-human agents.

By following a general trend, in the domain of digital humanities developers are progressively shifting from the project-driven approach to the new community-driven paradigm (Siemens *et al.* 2012). This shift is solicited by the increasing

¹ federico.boschetti@ilc.cnr.it.

² riccardo.delgratta@gmail.com.

³ angelo.delgrosso@ilc.cnr.it.

aggregation of scholars in communities that express common requirements and share best practices (Robinson 2013).

Along the decades, many initiatives have been financed to manage large varieties of documents (medieval manuscripts, handwritten contemporary documents, printed editions, etc., see Figure 1). Each project was focused on single philological aspects, *e.g.* the representation of variant readings and conjectural emendations, the diachronic development of author's variants, the treatment of critical apparatus (Bozzi 2014). The Text Encoding Initiative has established guidelines related to the formal representation of textual data for input, processing and output, (*cf.* again (McGann 2005)) but the agreement on interchange formats does not ensure interoperability. Indeed, as pointed out by Schmidt (2014), 'interchange (...) after a preliminary conversion of the data (...) implies some loss of information in the process', whereas 'interoperability' is 'the property of data that allows it to be loaded unmodified and fully used in a variety of software applications'.

In the community-driven paradigm, even the interoperability is not sufficient, because also the software components need to be reusable in different contexts and by different subcommunities (*e.g.* philologists and epigraphists in the wider digital humanities community).

For this reason, the role of each (sub)domain expert in the community is strategic. Their needs on topics such as content management, content searching/consuming are gathered through user-stories techniques and used for the overall design of components produced in the community.

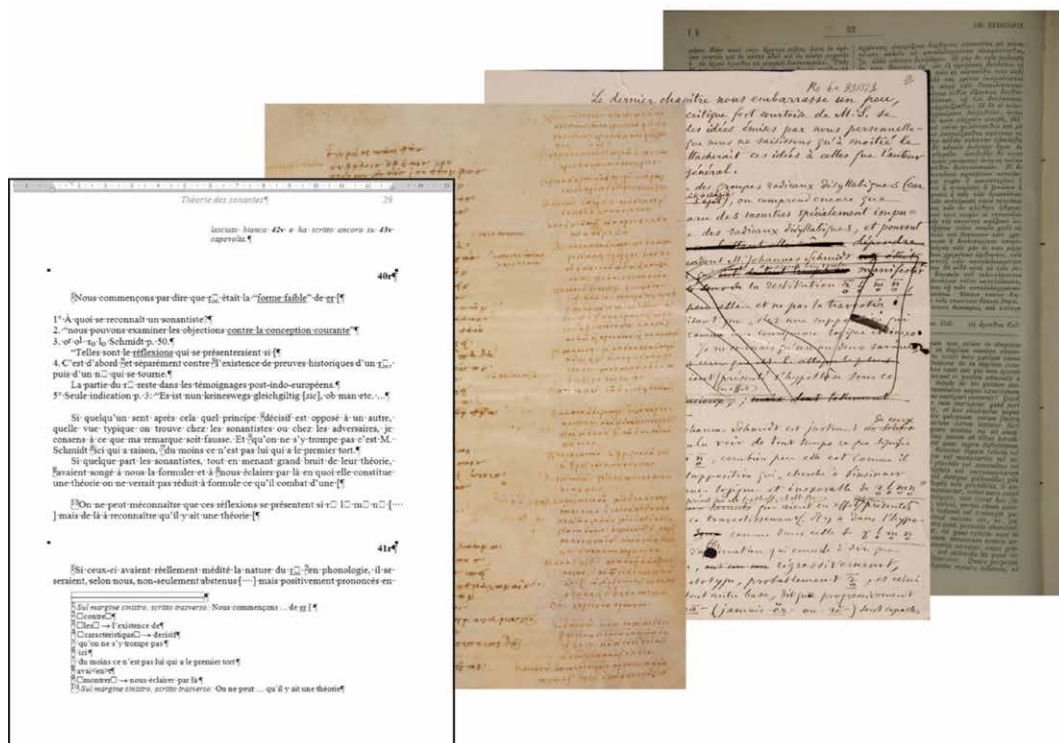


Figure 1: many documents and many kinds of philological issues.

In most cases, service providers are responding to these needs by offering web services quickly developed by taking into account the specific functionality that they expose or, worse, by wrapping legacy code. Although a pipeline of web services devoted to linguistic analysis and collaborative annotation provides many advantages in terms of flexibility, we are concerned by the impact of the main drawbacks, in order to study alternative or complementary solutions for our domain.

Modularity, maintainability, performance and atomicity are the principal issues in which we are interested.

Modularity is an architectural property that allows applications to be extensible, reusable, maintainable, and reactive to changes. The modularity of a single exposed service/application does not imply that each component of the service is modular in turn. For example a Greek part of speech tagger web service may be used in different contexts, but the single components of the service (e.g. the lemmatizer and the statistic model) could not be reused separately. This typically happens when web services are created by wrapping (pipelines of) legacy code.

Maintainability concerns what to do in case of service invocation failure⁴. In order to avoid failures, the system must be 'up-and-running' by assuring a reasonable level of service, which imply at least high-availability and recovery procedures. But small and medium size projects are not always able to assure such quality requirements. A second aspect that services need to address is related strongly to the typology and size of managed data.

Performance is affected by whether data or tools should be moved and by whether the service is called synchronously or asynchronously. Actually, performance is affected by the trade-off among challenging conditions (e.g. memory resources, computational overload, bandwidth). Defining such trade-offs is strongly related to the communities to whom the services are offered, since waiting one day for getting data might be reasonable for communities that manage large but static data, but also totally unacceptable for others that manage highly dynamic data. This is probably the most important reason for which the users are asked to provide a feedback.

Finally, atomicity assures that a service is capable of offering results in a consistent way, including rollback politics in case of failure. For example, if a user decides to remove an entire synset for the Ancient Greek WordNet (Bizzoni 2014), the service responsible for such operation needs to traverse and delete all the semantic and lexical relations which involve the synset and its lemmas or rollback to the initial status in case of failures.

At the Cooperative Philology Lab (Institute of Computational Linguistics 'A. Zampolli', CNR, Pisa) we aim to address these issues by designing and developing a library of components for the domain of scholarly editing (Robinson 2009). A library can be installed locally or remotely and can provide multiple choices for maintenance and performance tuning. But above all a library of components provides the building blocks to shape local or remote services at the adequate level of atomicity, in order to ensure reusability and extendibility.

4 URL unreachable for any reasons, timeout, maximum limit of size exceeded are the most common failures in web-services based architectures.

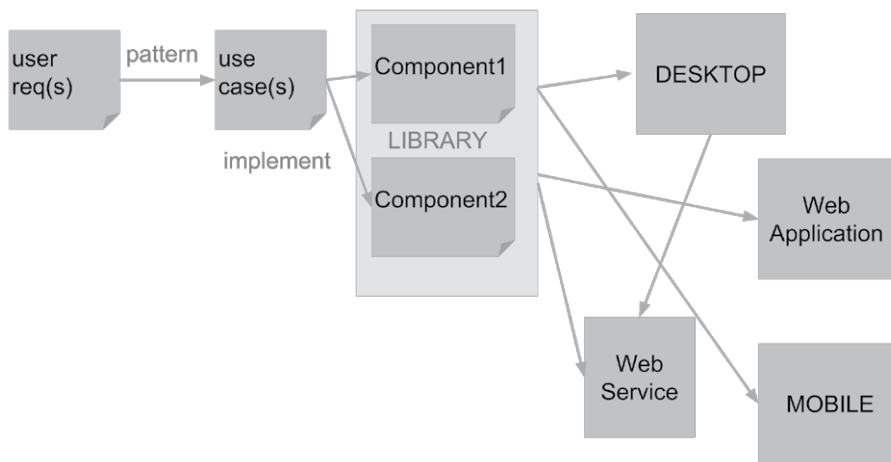


Figure 2: different applications built through the same library of components.

The role of the digital scholarly editors with which we have collaborated in pilot and funded projects at the CNR-ILC is crucial, because they are providing the necessary use cases that we are generalizing for the design of our library. In our experience, the multidisciplinary approach is enhanced when the team is lead by a digital and computational philologist with an interdisciplinary background. Indeed, the communication between humanists and software engineers is challenging, because for the domain experts it is difficult to linearly express their requests, if they are not able to recognize the technological limitations, and for the analysts it is difficult to elicit further information from them, in order to refine generic requests, if they are not able to master the peculiarities of the philological domain. But a new generation of digital humanists is emerging, and these scholars not only are creators of digital resources and consumers of computational tools or web infrastructures, but they also are actors in the analysis of requirements and in the evaluation of the computational instruments devoted to their activities.

In conclusion, in the age of software components and (web) services, the interaction between these two communities (DH and IT) can lead to a fruitful cross-fertilization, which aims at extending linguistic resources by textual evidence and enhancing scholarly editions by linguistic tools and lexico-semantic resources.

References

- Bizzoni, Y., F. Boschetti, H. Diakoff, R. Del Gratta, M. Monachini and G. Crane. 2014. 'The Making of Ancient Greek WordNet.' In *Proceedings of LREC 2014*.
- Bozzi, A. 2014. 'Computer-assisted scholarly editing of manuscript sources.' In *New publication cultures in the humanities: exploring the paradigm shift*, edited by P. Davidhazi. Amsterdam: Amsterdam University Press, 99-115.
- Kozar, O. 2010. 'Towards Better Group Work: Seeing the Difference between Cooperation and Collaboration.' *English Teaching Forum* 2.
- McGann, J. 2005. 'From text to work: Digital tools and the emergence of the social text.' *Variants: The Journal of the European Society for Textual Scholarship* 4: 225-240.
- Robinson, P. 2009. 'Towards a scholarly editing system for the next decades.' In *Sanskrit Computational Linguistics, Lecture Notes in Computer Science*, edited by G. Huet, A. Kulkarni, and P. Scharf. Berlin Heidelberg: Springer 5402, 346-357.
- Robinson, P. 2013. 'Towards a theory of digital editions.' *Variants* 10: 105-131.
- Schmidt, D. 2014. 'Towards an Interoperable Digital Scholarly Edition.' *jTEI* 7.
- Siemens, R., M. Timney, C. Leitch, C. Koolen, A. Garnett *et al.* 2012. 'Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media.' *Literary and Linguistic Computing* 27, no 4: 445-461.

Inventorying, transcribing, collating

Basic components of a virtual platform for scholarly editing, developed for the Historical-Critical Schnitzler Edition

*Stefan Büdenbender*¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

The *Trier Center for Digital Humanities* currently is involved in the creation of a number of digital editions, be it through preparing and providing data or through the developing of software solutions. In this context, the Arthur Schnitzler Edition (Arthur Schnitzler: Digitale historisch-kritische Edition (Werke 1905-1931)) holds a special position because of its complexity and scope.

The project was launched in 2012, with a projected runtime of 18 years. It is a binational cooperation, involving the Bergische Universität Wuppertal, the University of Cambridge, University College London and the University of Bristol, in partnership with the Cambridge University Library, the German Literary Archive at Marbach, and the Center for Digital Humanities at the University of Trier.²

It aims to create a critical edition of Schnitzler's works in digital form, to be published on an open access online platform. This portal is to bring together the physically dispersed archival holdings and the published works in a virtual form, combining the functions of digital archive and edition. The collected extant material – both manuscript and typescript – shall be digitally reproduced, transcribed, and made accessible through commentaries, registers etc.

With emphasis on the categories of 'textuality', 'materiality' and 'genesis', this results in a multitude of perspectives and views: the user can choose between diplomatic transcriptions, amended reading versions, and genetically interpreted reconstructions.

1 bued2101@uni-trier.de.

2 For more information, see: <http://www.arthur-schnitzler.de>.

To set the basis for this, two teams of philologists in Wuppertal and Cambridge have started to inventory and transcribe a corpus of more than 20,000 pages, documenting the textual development of all witnesses, down to the alteration of single characters, via a complex model of layers and states.

It became obvious in the early planning phase that there was no immediately available software solution that would cover the whole workflow. Given the amount of material to be edited and the long runtime, we decided to create a digital platform which is to support all associated steps of scholarly editing, from the first assessment and inventory of textual witnesses, through their transcription, down to the comparison of the resulting texts. The platform consists of individual modules and thus should be reusable by similar projects, be it in parts or as a whole. The main modules – some newly developed, others based on preexisting software – will be presented briefly below.

Technical Infrastructure: FuD

The research network and database system (*Forschungsnetzwerk und Datenbanksystem*) FuD³ forms the technical backbone of the platform, offering a set of features for decentralized collaborative work. On the one hand it allows the inventory (metadata capture, grouping) and commentary (creation of indexes) of the material, on the other hand it provides a database environment which manages all created content and thus represents the intersection between the individual modules. Although these have XML interfaces and can be run independently, the access to a central database facilitates collaboration and has advantages when dealing with concurrent hierarchies and structures transcending document borders.

Transcription of textual witnesses: Transcribo

Transcriptions are established in Transcribo⁴, a graphical editor developed to meet the exact needs of the project. Technically, it had to be able to communicate with FuD, to handle large image files without delay and to support an extensive set of specific annotations. The user interface sets the digital facsimile (generally the scanned physical witness) at the centre. This appears in double form, always providing the original for close examination while all processing steps take place on a slightly attenuated duplicate. This arrangement accommodates the use of multiple monitors and above all saves time-consuming jumping back and forth between image and editor window. Thus, field markings in rectangular or polygonal shape can be drawn around graphic units, reproducing their exact topography, and the transcribed text then can be entered directly. For the processing of typescripts an additional OCR with options for image enhancement (such as contrast and color adjustment) is integrated, providing a raw transcription as a basis for further editing.

3 <http://fud.uni-trier.de>.

4 <http://transcribo.org>.

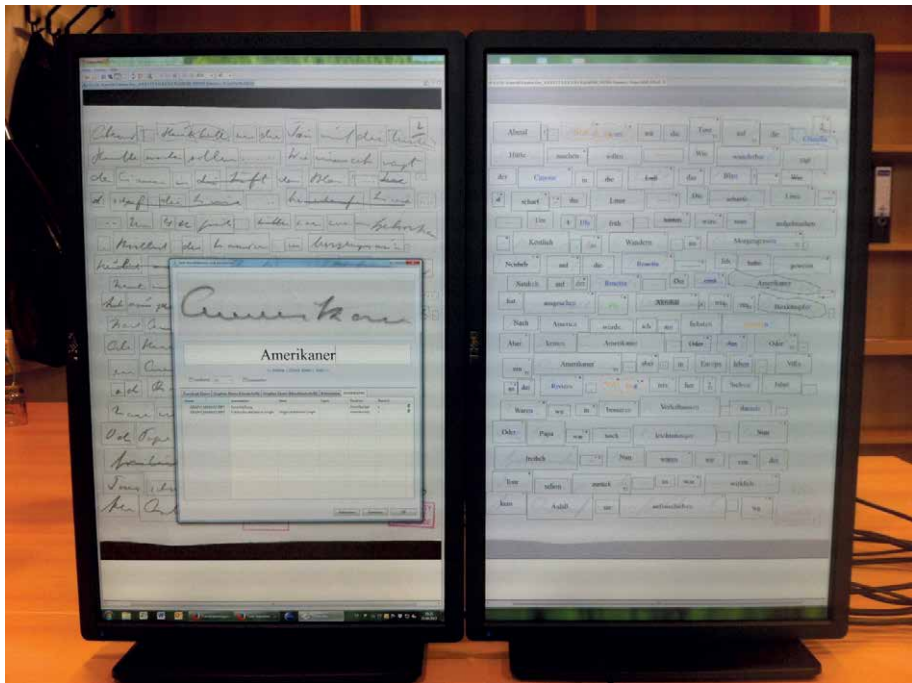


Figure 1: Using Transcribo in a dual-monitor setup.

More fundamentally, each transcribed element can be provided with a comment and each relevant philological or genetic phenomenon can be annotated in a uniform way. Here, a context menu is used with a set of project-specific options, with capacity for expansion. It will be adjusted as necessary throughout the project's lifecycle, according to the requirements of the textual basis.

In terms of data modelling, we found it very challenging to represent the results in a straightforward XML encoding, as there is a systematic overlap between the textual and material level of description, leading to the above-mentioned concurrent hierarchies for each and every text. While the TEI proposes a number of solutions to circumvent this inherent problem⁵, there are significant drawbacks to each of them when it comes to processing the data any further. Thus, the internal data exchange between the tools is handled via relational databases. When XML is required, be it for exchange with external partners or archiving, a TEI-conformant version can be created at any time.

Collation

Additionally, we are developing a graphic environment for textual comparison, as existing solutions have proven to be inadequate for a number of reasons. Firstly, the variation between the textual witnesses can at times be quite significant, as entire

⁵ As short overview is given in chapter 20 of the TEI guidelines: <http://www.tei-c.org/release/doc/tei-p5-doc/de/html/NH.html>.

paragraphs, chapters or scenes may be moved, split, joined, added or deleted. So, in many cases the challenge is to first spot the scattered counterparts before a meaningful comparison can be made. Secondly, it is often necessary to compare a large number of witnesses simultaneously. Finally, we want to visualise the results in different contexts, with varying granularity. The comparison process therefore is divided into two stages:

At first, all versions are put into chronological order and each one is compared to its successor in terms of larger linguistic or structural units (depending on the text: sentences, paragraphs, speeches, scenes etc.). These elements are then aligned, regardless of their textual position. Only after a complete matrix of corresponding passages through all witnesses has been established, checked by a philologist, and corrected if necessary, the matches are handed on for detailed comparison.

While this is a time-consuming approach, it has proven to work very well and it enables us to intervene at different stages of the collation. Moreover, we can evaluate different algorithms and existing solutions for the different tasks. Currently, we use parts of the TEI Comparator tool⁶ for the first alignment and CollateX⁷ for the actual comparison.

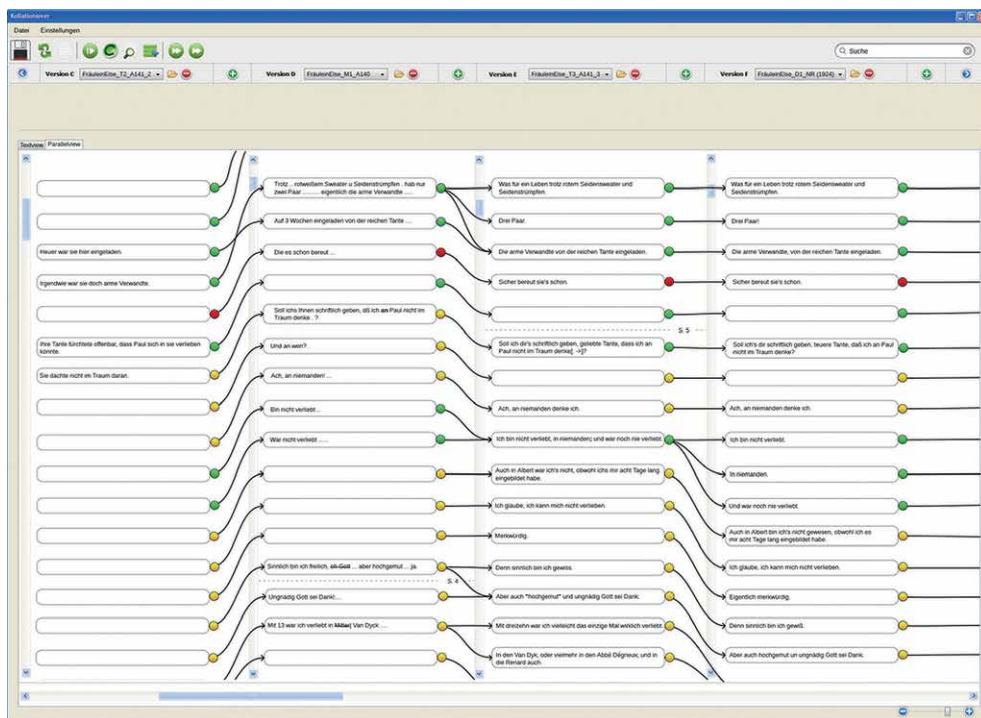


Figure 2: Building an alignment matrix for multiple textual witnesses.

6 <http://www.cems.ox.ac.uk/holinshed/about.shtml#tei>.

7 <http://collatex.net/>.

Outlook: material collection and genetic pathways

The platform is to be completed by an environment for ordering textual witnesses and defining genetic pathways. This is still in the design phase and should in due course make it possible to associate textual witnesses with the respective works or versions via drag and drop and to define sequences from various perspectives (genesis of a version, absolute chronological order, interdependence of version etc.).

Combining topic modeling and fuzzy matching techniques to build bridges between primary and secondary source materials

A test case from the King James Version Bible

Mathias Coeckelbergs,¹

Seth van Hooland² & Pierre Van Hecke³

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Living in a world where ever more documents are digitised, and every day many more born-digital are created, the demand for techniques to deal with this vast amount of data becomes ever more clear. Within the study of canonical historical texts, we see the publication of a large amount of new scholarly articles every year, that are not only important for researchers in related fields, but also for the digital critical edition of the primary source material. Hence the importance of inquiring how to produce meaningful links between primary and secondary source materials. In this article, we share our first experiments in this field.

The study of how texts are interlinked has been at the very core of the humanities ever since their inception. The scientific and technological developments resulted in a drastic increase of the scholarly production in the field. Seminal thinkers such as Paul Otlet (1934) and Vanevar Bush (1945) already developed the blueprints for how heterogenous documents could be interconnected in a meaningful way, but without having the technology at hand to implement the concepts and mental devices they imagined. Ted Nelson finally coined the term 'hyperlink' in the 1960s

1 mathias.coeckelbergs@ulb.ac.be.

2 svhoolan@ulb.ac.be.

3 pierre.vanhecke@kuleuven.be.

and presented it extensively in his work ‘Literary machines’ (Nelson 1980), in which he details the ambition to create a vast network of two-way connected nodes. Within Nelson’s vision, bidirectional links allow each node to understand where it is linked to and to analyse in detail the resulting network. However, hyperlinks as we all use and create them today in the context of the Web are unidirectional. The fact that everyone can create a link to a resource, without the requirement that the resource being linked to has to confirm the link and therefore remains agnostic of it, is central here. This was a conscious architectural decision from Tim Berners-Lee for the Web to scale rapidly in a radically decentralised approach.

The rise of the Semantic Web and the Linked Data paradigm have re-introduced the possibility of bi-directional relationships. Links between two resources are made through the use of an RDF (Resource Description Framework) triple. As the name suggests, it consists of three elements, respectively the subject, object and the predicate, which constitutes a relation between both. For example, if we have as subject ‘Guernica’, as predicate ‘painted by’ and as object ‘Picasso’, we have expressed in RDF that Picasso is the painter of Guernica. This method allows for easy encoding of many links for different resources, and makes it searchable by its own querying languages, SPARQL, allowing such queries as ‘what are all the resources painted by Picasso’ but also ‘give me all painters influenced by painters influenced by Picasso’, which are hard or impossible with standard online search engines. Although this method proves useful for linking multiple resources in a knowledge graph, it is not suitable for linking broader entities of linguistic meaning, such as sentences or paragraphs for example. The case study presented here will look closer at the methods and tools which can be used for the latter purpose.

In this paper we investigate the viability of combining two different methods of linking secondary scholarly work to the primary text, which in our case is the Bible. The first method is topic modeling, a technique aiming to extract keywords from a given collection of documents, which are afterwards clustered using a probabilistic algorithm. Throughout the past 25 years, many variations have been proposed to two main types of topic modeling algorithms, i.e. Latent Semantic Analysis (LSA – based on vectorial representation of words/documents; Deerwester *et al.* 1990) and Latent Dirichlet Allocation (LDA – working with prior probabilities; Blei *et al.* 2003). Since it is by far the most commonly used today, and because space does not permit us to compare techniques, we have opted to choose LDA for our experiments. The second method we consider is a fuzzy matching algorithm to link biblical verses to journal articles that cite them. After we have discovered the possibilities and limits of both approaches, we will describe in the second section how a combination of both is able to provide a best of both worlds solution.

Comparison of Both Techniques and Evaluation of Results

Topic Modeling

In this article, we apply LDA to both the King James Version of the Bible, as well as to 76,204 scholarly articles on the bible extracted from JSTOR digital library, satisfying the query 'bible'. Topic modeling algorithms start with the raw text, on the one hand, and the user's decision of how many topics need to be discerned in the text. Based on this initial input, the Latent Dirichlet Allocation algorithm, which we use in this article, generates word distributions per topic, and topic distributions per document, respectively. Informed by the Dirichlet priors, indicating the prior probability values, and iterating this process of sampling updated probabilities for both distributions via Gibbs sampling, we end up with a clustered groups of keywords extracted from the text. These clusters then constitute the computational topics, which the model assumes to be latently present explaining the present distribution of vocabulary words. Hence, we can infer from a list of keywords *Israel land David king Solomon* that we deal with the topic of kings of Israel. It is important to note, however, that the algorithm only presents the clustered keywords, and that the inference to the label of the topic is made by the researcher.

As is evident from this short overview of the LDA methodology, moving from a computational topic to a concept as understood by humans, requires interpretation. In addition, it is equally challenging to estimate how to objectively compare the results of different topic models. What is the influence of extracting amounts of topics which lie close to each other? How does deleting or adding a text to the corpus affect the extracted topics? These are only two questions to which a satisfying answer is hard to find. Space does not permit us to discuss in detail a comparison of the extracted topics. Suffice it to say that the most stable clusters of identifiable semantic units are obtained when between 300 and 400 topics are extracted. Below, we will see how we can delimit the interpretation radius of the extracted clusters.

Up to this point, we have only placed emphasis on the extraction of semantics from the text, not on how it actually constitutes links between texts. The algorithm presents us with a collection of topics which are presumed to be latently present within the collection of documents. For each topic, we have information on how the words from the corpus and all documents relate to it. Links between documents can be estimated from their similarity in ranking for the same topics. For example, if we find two scholarly articles dealing with the creation theology in the book of Job, it is likely that several key words from one article also will appear frequently in the other, resulting in them having highly similar scores for the topic of creation.

In this way, we can construe a comparison of scholarly articles dealing with the Bible, without referencing the source text. If we want to compare the extracted clusters from the articles on the one hand with those extracted from the Bible on the other, the issue of topic comparison we described above arises again. Computational topics can be compared by their similarity in keywords, but it would be risky to conclude that scholarly articles having high correspondence to one topic would be relevant for biblical documents scoring above threshold

similarity with a similar topic extracted from the bible. Another possibility would be to establish a topic model for both the biblical text as well as scholarly articles simultaneously. Although this approach will provide clusters of keywords to which both biblical and scholarly documents are linked, these models will for the greatest part report links between scholarly articles, because they are far greater in number. Hence, we can conclude that both approaches have their shortcomings, and that with topic modeling used the way we proposed here, it is hard to make a comparison between the original primary text and secondary literature. From this we conclude that, although topic modeling provides salient results in classifying documents according to relevance for a certain cluster of keywords, it is hard to use this technique for linking primary and secondary texts.

Fuzzy Matching

In contrast to the previous subsection, which focused on inferring more abstract units of semantics based solely on the input of raw text, we now focus on individual sentences appearing in the corpus. As with any task of matching elements from one set to another, possibilities of incorrect links are possible. In this case, we talk both about false positives (established links which are incorrect) as well as true negatives (quotations which are not linked to an article in which they appear).

The evaluation of the results of fuzzy matching is conceptually easier than the method of topic modeling described above. We do not have the space here to describe the exact results of different levels of similarity score in detail, but we have taken over the demands for declaring matches from the general description of the matchmaker API developed by JSTOR labs. These requirements are that similarity is above 80%, and when at least fifteen characters match, results are considered satisfactory.

Different degrees of uncertainty can occur concerning links, which human interpreters have no difficulty in mapping to the same entity. This is an important difference with topic modeling, where the interpretation of a computational topic to a human concept can lead to divergent outcomes among human annotators. For our fuzzy matching approach, a simple example of multiple entities with the same referent is the reference of the verse identifier. Many versions are thinkable. First of all, the book name can appear fully, such as for example ‘Genesis’, but also in abbreviated form, as ‘Gen.’ or ‘Gn’. Of course, it is important that all these instances are mapped to the same verse. Secondly, the verse numbers can be added and separated by different signs, of which commas and semicolons are used the most often.

Although for our work on topic modeling we ran experiments on both the Hebrew original as well as the King James Version (the most widely used English translation of the Bible), we have opted to limit our research to English examples only, for the simple reason that most scholarly articles are written in English. Of course, in the future we wish to include other languages as well (and experiment equally with the linking of Hebrew articles to the Hebrew original version). For the moment we run into some difficult problems regarding multilingual topics, to which we come back in section 2.2, that deals with the persisting difficulty of dealing with a multilingual context for both approaches.

An evident drawback of the fuzzy matching approach is that during the linking process it only matches quotations to scholarly articles. Hence, it does not make a classification of the latter articles in any way, apart from indicating which articles are citing which verses, and which articles contain how many quotations of biblical verses. This leaves us with the problem that, certainly for often-cited verses, the list of linked articles is hard to oversee without manually reading through all of them.

Possibility of Mutual Enrichment

As we have concluded in the previous section, both the approach of topic modeling as well as that of fuzzy matching come with their own advantages, but also disadvantages. By bringing both approaches together, however, the potential of both can be increased. We are still in the process of working out the methodology proposed here. Hence, in this contribution we will only describe the plan for conducting this research.

Since the fuzzy matching approach presents highest scores on correctness, it is taken as the point of departure and its matching results are improved with topic modeling, rather than the other way around. In this way, topic modeling presents a tool to be able to classify the results of fuzzy matching links to a sentence. Also, the major drawback of fuzzy matching is that we are left with unsorted articles linked to individual verses, a sorting which can be done by topic modeling. On the other hand, the problem of interpretation of extracted topics can be leveraged by constraining the interpretation radius by verse citation. Since the fuzzy matching approach links verses to articles citing them, the links will per definition be meaningful, but the extent remains to be discovered by topic modeling.

Proof of concept of combining Fuzzy Matching and Topic Modeling

The first step in our methodology for combining the strengths of both topic modeling and fuzzy matching is to model topics from the scholarly articles found in the JSTOR database, which in our previous approach already were linked to bible verses using ‘Matchmaker’, the fuzzy matching algorithm developed by JSTOR Labs. We propose to manually label the extracted topics, so that general, human-made topics are available based on the computationally generated ones.

Taking from the first section our approach of fuzzy matching, we end up with on the one hand all biblical verses, linked to a set of scholarly articles in which the verse under consideration is quoted. On the other hand, we have the computational topics generated on the basis of the entire collection of these articles, and human labels to make them easier to interpret. Our second step, then, renders visible for a selected verse not only the articles it is linked to, but also the entire list of topics, and the score for their recurrence in the linked articles. In this way, the user can easily derive in which context the verse under scrutiny recurs, without having to read through all the linked articles.

What would of course be most valuable is that the algorithm automatically selects articles relevant to a user query. Query expansion techniques present possibilities here, although we believe for the current goal of structuring articles linked to a specific verse, it might be easier to present the user with a set of key

terms for which the resulting articles can be sorted by manually labeling the computational topics extracted from the corpus of scholarly work, the inferred labels can be used as key terms to appear after a verse has been selected by selecting the topic the user is interested in, he can see for a verse he is investigating which articles are most relevant. On the other hand, it is straightforward that for the results we already have generated, it is equally possible for a user to select a key term, find which documents are most relevant for these topics, and hence see which verses are most popular in being quoted in this context.

The problem of multilingualism

A persisting problem we have encountered throughout our research is that of multilingualism. Not only is the primary source text written in dead languages (Classical Hebrew, Aramaic and Greek), the secondary sources appear in numerous languages as well. For the present purpose of showing the viability of our approach, we have limited our choice of languages to English, although salient topic models also have been developed for Hebrew. Because the English collection is understandably bigger than the Hebrew one, we selected the one with the highest resources. Of course, we want to be able to go further than to link primary text or its translation only to documents that happen to be in the same language.

A solution to this problem will consist either of developing polylingual topic models, or by ‘translating’ the extracted topics to each other, or to interlingual topics. Concerning the first type of models, most research has been conducted in context of high topical similarity, for example in Wikipedia articles across several languages dealing with the same issues. Although these models provide promising results, our research poses a more difficult problem in sorting all articles in different languages according to a mutual set of topics. More research is needed in this respect. A viable way out seems to us to define a topic ‘translation’ model, which classifies two topics in different languages as equal topics if and only if a threshold of dictionary nearness has been reached for most of the words in both topics.

Conclusion

The goal of this article was to show that combining Fuzzy Matching and Topic Modeling techniques is a viable method for generating links between primary and secondary source materials. We have seen that both methods in themselves have their assets and drawbacks, but when combined a satisfying solution can be found, which will be worked out further in the future.

References

- Blei, David M., Andrew Y Ng. and Michael I. Jordan. 2003. 'Latent Dirichlet Allocation.' *Journal of Machine Learning Research* 1.3: 993-1022.
- Bush, Vannevar. 1945. 'As we may think.' *The Atlantic Monthly* 176.1: 101-108.
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. 'Indexing by Latent Semantic Analysis.' *Journal of the American Society for Information Science* 41. 96: 391-407.
- Nelson, Theodor H. 1980. *Literary Machines*. Sausalito, CA: Mindful Press.
- Otlet, Paul. 1934. *Traité De Documentation: Le Livre Sur Le Livre, Théorie Et Pratique*. Bruxelles: Editiones Mundaneum.

The importance of being... object-oriented

Old means for new perspectives in digital textual scholarship

Angelo Mario Del Grosso,¹

Emiliano Giovannetti² & Simone Marchi³

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Introduction

In this contribution we propose an Object-Oriented (OO) approach to the design and development of reusable tools for the domain of scholarly editing. It regards some software engineering considerations about the importance of being object-oriented in implementing software applications for digital textual scholarship. This work fits into an ongoing discussion about textual modelling (Pierazzo 2015) where the need for extensible, reusable and modular applications is constantly increasing (Driscoll and Pierazzo 2016).

Although the digital turn of textual scholarship is nowadays a reality and many advancements have been made in encoding and visualizing textual resources, flexible and shared models in the construction of tools for scholarly editing are still missing (Almas and Beaulieu 2013; Shillingsburg 2015; Robinson and Bordalejo 2016). This lack typically leads to the development of *ad hoc* – i.e. not reusable – software (Ciotti 2014; Schmidt 2014).

1 angelo.delgrosso@ilc.cnr.it.

2 emiliano.giovannetti@ilc.cnr.it.

3 simone.marchi@ilc.cnr.it.

Exploiting our experience in projects we worked on (e.g. Giovannetti *et al.* 2016; Abrate *et al.* 2014; Bozzi 2013; Del Grosso *et al.* 2013), we have conceived a general OO model that will be introduced in the following section (Del Grosso *et al.* 2016; Boschetti and Del Grosso 2015).

Method and Discussion

The typical approach to the digital scholarly editing generally encompasses two phases: 1) textual encoding and 2) web publishing. On the one hand, the data and metadata encoding phase makes the textual resource machine-actionable and useful for information exchange. The best practice includes the identification of textual phenomena by marking up the transcribed source by means of suitable XML tags following vocabularies which are defined in specific and formal schemas (e.g. the TEI-XML guidelines). On the other hand, the aim of the electronic publication phase is to provide end-users (mainly scholars) with high-level functionalities (e.g. advanced visualization and searching) through Web graphical user interfaces. Actually, powerful frameworks built on Javascript libraries – such as AngularJS, Ember, React or D3.js – provide developers with ready-made widgets useful for the processing and publishing of digital editions. This approach can be adopted just in large academic project-oriented initiative, because, in our opinion, it requires too much developing work in order to implement and/or customize the specific digital environment.

In addition, this strategy also brings some negative consequences from an engineering point of view. Indeed, it lacks of (1) formal specifications for the domain of interest; (2) abstractions for shared models; and (3) plans for software evolution and software reuse.

The lacks concerning data and procedure abstraction become particularly evident when software designers and developers strive to encapsulate both the digital representation of the textual resource and the operations required to process it. This matter is even more challenging when software architects look for formal and shared models to develop reusable components. In the light of all of this, we are working on the definition of some textual scholarship entities (e.g. the document, the text, the edition, the witness, etc.) together with the functions needed to manipulate them as shared and implementation-independent objects. Figure 1 shows a diagram of our Object-Oriented model for textual scholarship designed starting from the identification of the Domain Specific Abstract Data Types (DS-ADTs).

An ADT is a high-level and mathematically-founded data type the internal representation of which is not directly accessible from users (information hiding). Actually, data and functions which operate on them are bound into a single entity of concerns.

By adopting this approach, the focus turns from the data value and representation towards the behavior of the components. Moreover, suitable Application Programming Interfaces (APIs) define functionalities and protocols to easily integrate, extend and use software components in different projects. The proper design of APIs is a critical task within Domain Driven Applications

(Bloch 2006), since they are the only point of dependence between clients (users) and service providers (namely, who implements the ADT). Finally, within our approach, we also apply Design Patterns as ‘off-the-shelf’ solutions to recurring problems in developing software for specific contexts (Ackerman and Gonzalez 2011). Figure 2 shows an UML diagram of the textual analysis component embodying the aforementioned design policies.

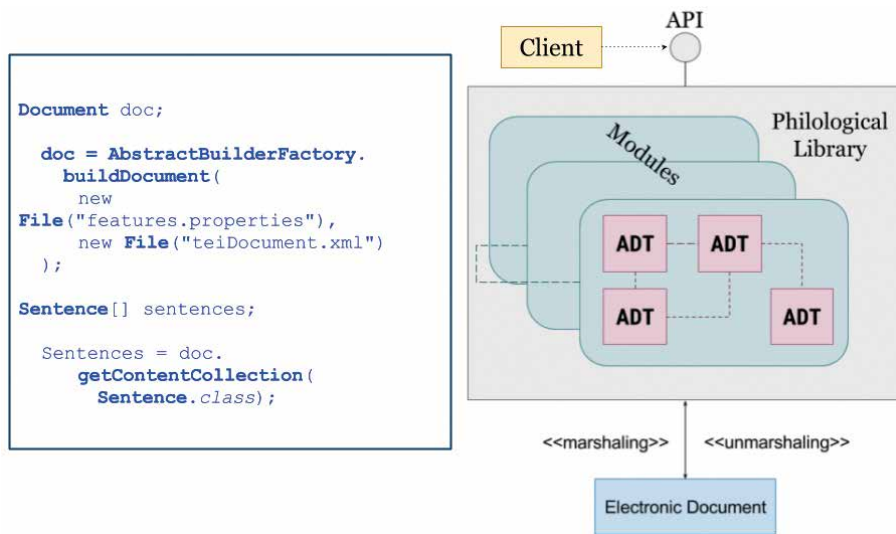


Figure 1: The Domain Specific Abstract Data Type approach (DS-ADT): (on the left) **doc** is an instance of the ADT **Document Class** representing the encoded document. The DS-ADT hides the implementation details and exposes all and only the necessary functionalities.

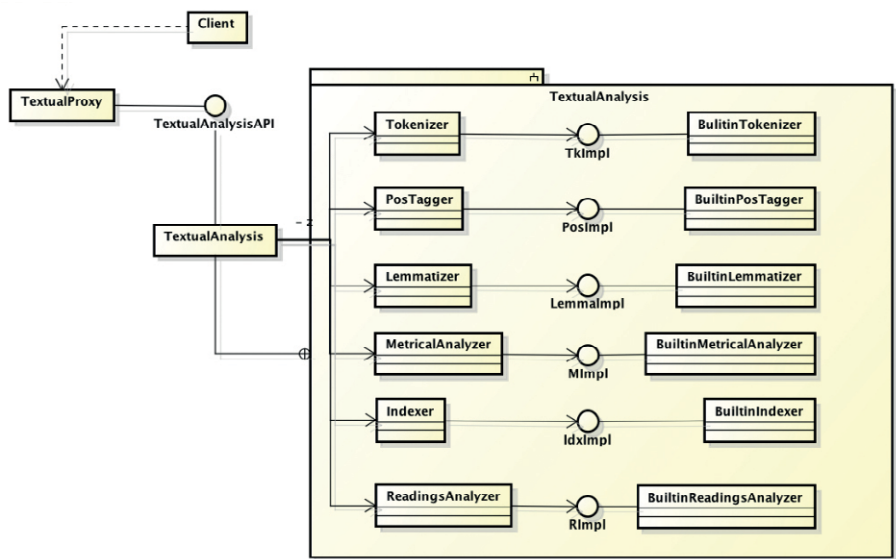


Figure 2: Example of the object-oriented design principles. Such a design provides an effective way to implement decoupled, abstract and reusable software components.

Once the architecture, the model and the ADTs have been defined, it is necessary to put in place a process to implement the tools. To do this, we chose, as our development environment, the Object-Oriented Analysis, Design and Development – inspired from the ICONIX approach (Collins-Cope *et al.* 2005). In particular, our process is structured in five main steps, roughly corresponding to the typical software engineering workflow: A) the involvement of the scholar community (Cohn 2004) to gather the requirements and define the specifications (Rosenberg and Stephens 2007); B) the design of the single components following the Domain-Driven and the User-Centered approach (Evans 2014); C) the design of the general architecture using the Pattern-Oriented approach and the UML diagrams (Fowler 2003); D) the development of the software according to the S.O.L.I.D. principles;⁴ and E) the implementation of advanced Graphical User Interfaces through a specific UI framework.

Conclusion

From a software engineering perspective, digital textual scholarship in general, and digital scholarly editing in particular, need a formal definition of digital domain entities and procedures: this formalization is mandatory to design and develop of what we call the Domain Specific Abstract Data Types for Digital Scholarly Editing (DS-ADTs for DSE).

These ADTs will lead, then, to the identification of standard and shared Application Programming Interfaces (APIs) and to the specification of the components' behavior. The APIs are, by definition, independent both from the actual representation of the data and the algorithms manipulating them. In this way, we ensure a strong decoupling between the API user and the API developer. Moreover, the OO approach can bring benefits not only to developers but also to end users (as scholars) who could count on a more efficient and effective development of the tools they need.

From a practical point of view, we are applying the principles illustrated in this contribution by implementing a digital scholarly platform, called Omega, which is hosted on github (<https://github.com/literarycomputinglab>).

⁴ Single responsibility, Open-closed, Liskov substitution, Interface segregation, Dependency inversion

References

- Abrate, Matteo, Angelo Mario Del Grosso, Emiliano Giovannetti, Angelica Lo Duca, Damiana Luzzi, Lorenzo Mancini, Andrea Marchetti, Irene Pedretti, and Silvia Piccini. 2014. 'Sharing Cultural Heritage: The Clavius on the Web Project.' In *Proceedings of the 9th LREC Conference*, Reykjavik, 627-634. ELRA.
- Ackerman, Lee, and Celso Gonzalez. 2011. *Patterns-Based Engineering: Successfully Delivering Solutions Via Patterns*. Addison-Wesley.
- Almas, Bridget, and Marie-Claire Beaulieu. 2013. 'Developing a New Integrated Editing Platform for Source Documents in Classics.' *LLC* 28. 4: 493-503.
- Bloch, Joshua. 2006. 'How to Design a Good API and Why It Matters.' In *Companion to the 21st ACM SIGPLAN Symposium (OOPSLA)*, Portland, Oregon, USA, 506-507.
- Boschetti, Federico, and Angelo Mario Del Grosso. 2015. 'TeiCoPhiLib: A Library of Components for the Domain of Collaborative Philology.' *JTEI Journal* 8.
- Bozzi, Andrea. 2013. 'G2A: A Web Application to Study, Annotate and Scholarly Edit Ancient Texts and Their Aligned Translations.' *Studia Graeco-Arabica* 3: 159-171.
- Ciotti, Fabio. 2014. 'Digital Literary and Cultural Studies: State of the Art and Perspectives.' *Between* 4. 8.
- Cohn, Mike. 2004. *User Stories Applied: For Agile Software Development*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc.
- Collins-Cope, Mark, Doug Rosenberg, and Matt Stephens. 2005. *Agile Development with ICONIX Process: People, Process, and Pragmatism*. Berkely, CA, USA: Apress.
- Del Grosso, Angelo Mario, Davide Albanesi, Emiliano Giovannetti, and Simone Marchi. 2016. 'Defining the Core Entities of an Environment for Textual Processing in Literary Computing.' In *DH2016 Conference*. 771-775. Kraków: Jagiellonian University and Pedagogical University.
- Del Grosso, Angelo Mario, Simone Marchi, Francesca Murano, and Luca Pesini. 2013. 'A Collaborative Tool for Philological Research: Experiments on Ferdinand de Saussure's Manuscripts.' In *2nd AIUCD Conference*. 163-175. Padova: CLEUP.
- Driscoll, Matthew James, and Elena Pierazzo (eds). 2016. *Digital Scholarly Editing: Theories and Practices*. Vol. 4. Digital Humanities Series. Open Book Publishers.
- Evans, Eric. 2014. *Domain-Driven Design Reference: Definitions and Pattern Summaries*. Dog Ear Publishing.
- Fowler, Martin. 2003. *UML Distilled: A Brief Guide to the Standard Object Modeling Language*. 3rd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Giovannetti, Emiliano, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2016. 'Traduco: A Collaborative Web-Based CAT Environment for the Interpretation and Translation of Texts.' *Digital Scholarship in the Humanities*.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods. Digital Research in the Arts and Humanities*. Farnham Surrey: Ashgate.

- Robinson, Peter, and Barbara Bordalejo. 2016. 'Textual Communities.' In *DH2016 Conference*, 876-877. Kraków: Jagiellonian University and Pedagogical University.
- Schmidt, Desmond. 2014. 'Towards an Interoperable Digital Scholarly Edition.' *JTEI Journal*, no. 7.
- Shillingsburg, Peter. 2015. 'Development Principles for Virtual Archives and Editions.' *Variants* 11: 9-28.
- Rosenberg, Doug, and Matt Stephens. 2007. *Use Case Driven Object Modeling with UML: Theory and Practice*. (New ed.). Berkeley, Calif.: Apress.

Edition Visualization Technology 2.0

Affordable DSE publishing, support for critical editions, and more

Chiara Di Pietro¹ & Roberto Rosselli Del Turco²

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The popularity of digital scholarly editions has grown considerably during the last decade, at least judging by the number of projects completed or nearing completion. Many of the existing editions, however, are the result of sophisticated programming and implementation of complex, feature-rich frameworks: while text encoding is relatively easy, a single scholar or a small group of researchers surely would have to look for further support and resources to publish the resulting edition.

Edition Visualization Technology (EVT)³, an open source tool to produce digital editions on the basis of XML TEI-encoded documents, was born to serve the goals of a single project, the Digital Vercelli Book (<http://vbd.humnet.unipi.it/beta2/>), but has been developed in such a way as to become a general purpose tool. This software aims at freeing the scholar from the burden of web programming, so that (s)he can focus on preparing the edition documents according to the TEI Guidelines and schemas. After the web edition is generated using EVT, the final user can browse, explore and study it by means of a user-friendly interface, providing a set of tools (zoom, magnifier and hot-spots for manuscript images, text-image linking and an internal search engine for the edited texts) for research purposes.

1 dipi.chiara@gmail.com.

2 roberto.rosselidelturco@unito.it.

3 For more information on the EVT, see: <http://evt.labcd.unipi.it/>. For the EVT's file repository, see: <https://sourceforge.net/projects/evt-project/>. For the EVT's code repository, see: <https://github.com/evt-project/evt-viewer>.

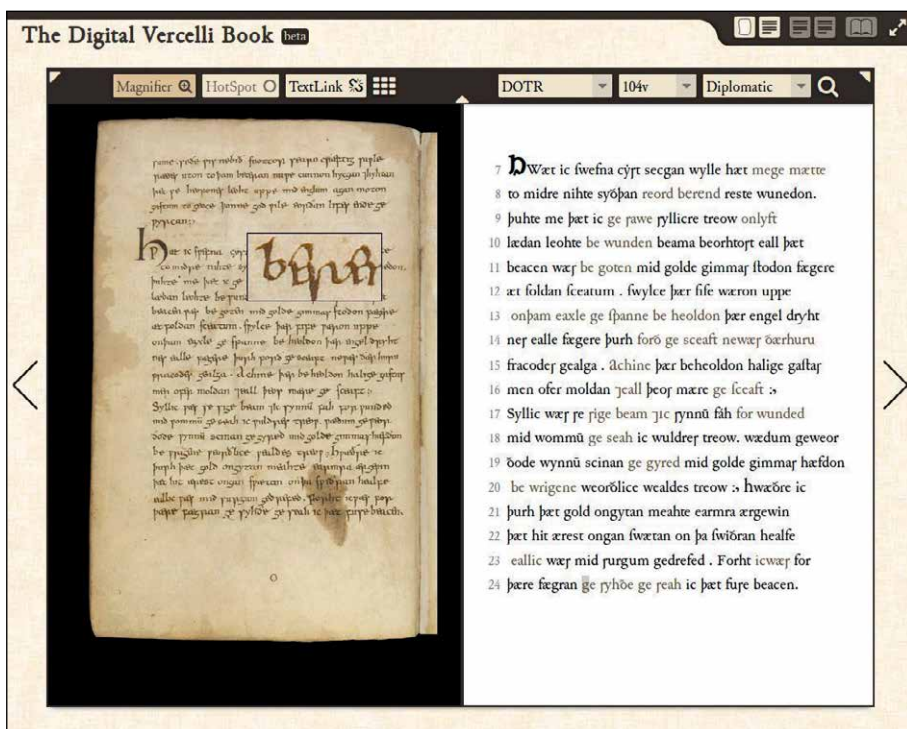


Figure 1: The first beta version of the Digital Vercelli Book (2014).

The starting point of the system is one or more documents in the standard TEI P5 format: by applying a single XSLT style-sheet, the TEI XML text is turned into a web based application – a mix of HTML5, CSS3 and JavaScript – that can be easily uploaded and shared on the Web; in fact, since the EVT viewer is based on the client-only model, there is no need to install and configure additional software on the server. The text can be presented in different levels of edition (e.g. interpretative, diplomatic) and, besides the default visualization layout where text and scans of the original manuscript are linked together and placed side by side, a book reader mode can be activated if double side images are supplied.⁴

Version 1.0 has been released in February 2016 and has proved to be quite successful, since it has been adopted by many other projects such as the *Codice Pelavicino Digitale* (<http://labcd.humnet.unipi.it/evt/>) and the *Tarsian Project* (<http://tarsian.vital-it.ch/about/>). There are also many new projects led by young researchers who have found in EVT the perfect tool for their needs, some of these will be announced during the months to come.

As a consequence of these collaborations, EVT has been enriched with several new features, and – being open source software – each new feature added as per request by a specific edition project is going to benefit all others (if applicable). The continuous development and need to adapt EVT to different types of documents and TEI-encoded texts has shifted the development focus towards the creation of a more flexible tool for the web publication of TEI-based documents, able to cater to multiple use cases. One of the most requested features, and actually one of the Digital Vercelli Book project's original requirements, is the support not only for diplomatic transcriptions linked to the corresponding manuscript images, but also for critical editions complete with a proper apparatus. The complexity of this task and of other planned additions, combined with the growing intricacy of the current code base – especially with regard to adding and combining new features – convinced the development team that it was essential to move over and start with a complete code rewrite for the following version.

The current version (EVT 1), therefore, still is being developed and will be supported with bug fixes / small enhancements for a long time, but it will be the last one to use the current, XSLT-based architecture. EVT 2.0 is already under development using a different approach in order to improve flexibility and modularity, to make it easier to implement new features and to adjust the UI layout for different kinds of editions. This is why the development team decided to refactor the whole code of the viewer and base it on AngularJS (<https://angularjs.org>), a Javascript framework which implements the MVC (Model View Controller)⁵ pattern to separate presentation, data and logic components, providing a great modularity of the web application. The goal is to offer a tool that can be customized easily and does not need any intermediate XSLT transformations. The user just needs to point at the encoded file by setting its URL in a configuration file, and to open the main page: the web-edition will be automatically built upon the provided data. As the previous version, EVT 2 is a client-only tool, with no specific server required, the user will be able to deploy the web-edition immediately on the Web.

⁴ See Rosselli Del Turco 2015 for more information.

⁵ See: <https://en.wikipedia.org/wiki/Model-view-controller>.

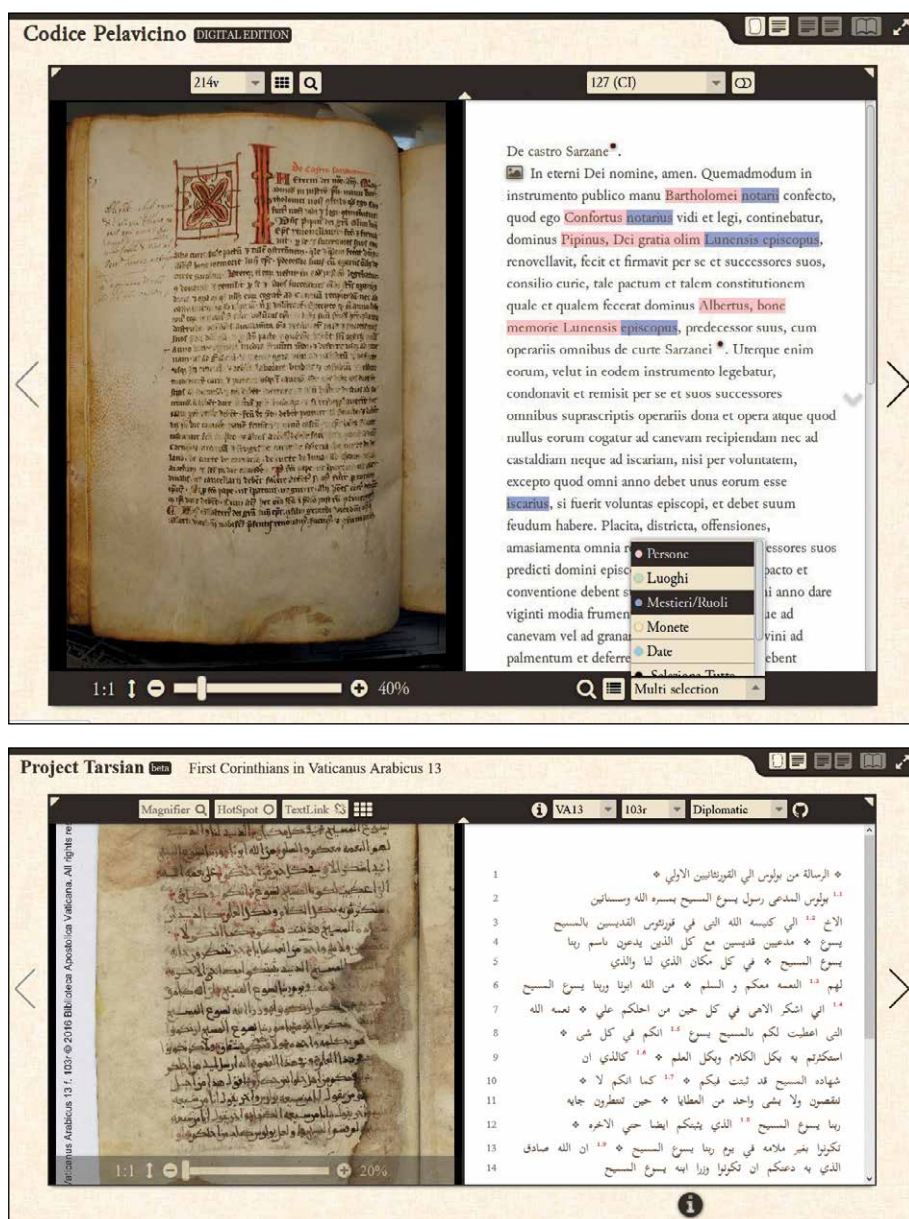
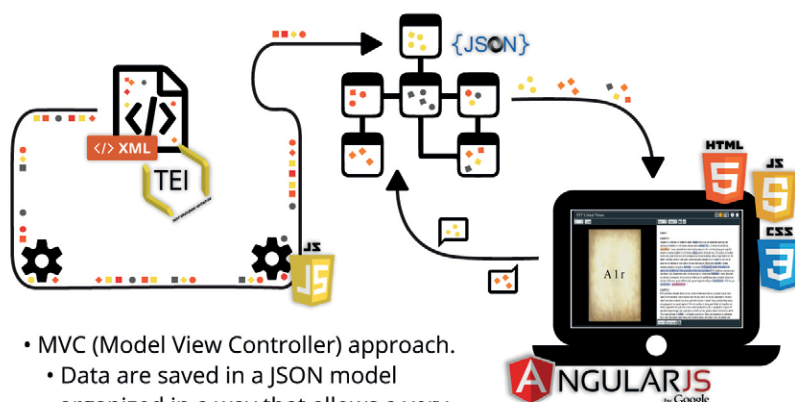


Figure 2: The Codice Pelavicino Digitale (top) and the Tarsian digital edition (bottom).

The current development focus revolves around the support for critical editions, which is a very challenging goal in itself, especially with regard to the complexity of the User Interface layout (Rosselli Del Turco 2011). This new level of edition is based on the current TEI relevant CA module and Guidelines chapter and it supports the Parallel Segmentation method. The current implementation, however, is meant to be as generic and flexible as possible to make it easier to update it when the TEI module will be rewritten and expanded to become more powerful and suitable to philologists.



- MVC (Model View Controller) approach.
 - Data are saved in a JSON model organized in a way that allows a very quick access to the information needed.
- No more XSLT transformations.
- Direct parsing of the XML.

Figure 3: The new architecture for EVT 2.0.

Doc 1

Critical

Liber I

CAPUT 1

Magnus es, domine, et laudabilis valde: magna virtus tua, et sapientiae tuae non est numerus et laudare te vult homo, aliqua portio creature tue

creature tue]	creatr ae tuae	A creaturastuas	B creaturarum tuarum	E C D
---------------	----------------	-----------------	----------------------	-------

Critical Note
More Info
XML

, et homo circumferens mortalitem • suam, circumferens testimonium peccati sui et testimonium, quia superbis resistis: et tamen laudare te vult homo, aliqua portio creaturae tuae. Tu excitas, ut laudare te delectet, quia fecisti nos ad te et inquietum est cor nostrum, donec requiescat in te. Da mihi • domine, scire et intellegere, utrum sit prius invocare te an laudare te, et scire te prius sit an invocare te. sed quis te invocat nesciens te? Aliud enim pro alio • potest invocare nesciens. An potius invocaris • ut sciaris? In Quomodo autem invocabunt, in quem non crediderunt? Aut quomodo credent sine praedicante? Et laudabunt dominum qui requirunt eum. Quaerentes enim inveniunt eum et invenientes laudabunt • eum. Quaeram te, domine, invocans te, et invocem te credens in te: praedicatus enim es nobis. Invocat te, domine, fides mea, quam dedisti mihi, quam inspirasti mihi per humanitatem • filii tui, per ministerium • praedicatoris tui . •

CAPUT 2

• Et quomodo invocabo deum meum, deum et dominum meum,

Filters
Heat Map
A

Figure 4: EVT 2: base text with inline critical apparatus.



Figure 5: EVT 2: base text collated with witnesses.

Among the different tools offered, EVT 2 provides a straight and quick link from the critical apparatus to the textual context of a specific reading; moreover, it allows for comparing witnesses' texts among each other or with respect to the base text of the edition (or to another specific text); finally, it offers specific tools such as filters, to select and show different textual phenomena (e.g. the person responsible for each emendation), and a heat map, showing where the variants are more frequent in the textual tradition. All these features are implemented already and can be tested by downloading the current development version from the file repository or straight from the code repository. In the final version, the user will be able to examine each variant in its paleographic context if the digitized images of each manuscript are provided.

From the point of view of the editor, the new architecture will be as easy to use as the current one: the only technical skill required will be a general competence in XML editing in order to configure EVT properly and to place each XML-related component of the edition (mainly the schema besides the encoded texts) into the correct area of the directory structure. For those editors who can boast computer programming skills, or who can count on technical support, there also will be the possibility to add new CSS rules and to customize all aspects of text visualization according to their needs.

In conclusion, EVT 2 is going to be a greatly enhanced version of an already popular tool enabling the single scholar, or a small research group, to publish digital editions, both diplomatic and critical, in an easy and effective way.

References

- Rosselli Del Turco, Roberto. 2011. 'After the Editing Is Done: Designing a Graphic User Interface for Digital Editions.' *Digital Medievalist* 7. <http://www.digital-medievalist.org/journal/7/rosselliDelTurco/>.
- Rosselli Del Turco, Roberto, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti, and Jacopo Pugliese. 2015. 'Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions.' *Journal of the Text Encoding Initiative* 8. DOI:10.4000/jtei.1077. <http://jtei.revues.org/1077>.
- TEI Consortium (eds), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. (3.0.0). (2016-03-29). TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>.

Compilation, transcription, multi-level annotation and gender-oriented analysis of a historical text corpus

Early Modern Ducal Correspondences in Central Germany

Vera Faßhauer¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Introduction

The corpus *Early Modern Ducal Correspondences in Central Germany* (*Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum*) consists of 600 Early New High German letters comprising 262,468 tokens. It was created between 2010 and 2013 during a DFG-funded research project located at the University of Jena and carried out by Rosemarie Lühr, Vera Fasshauer, Henry Seidel and Daniela Prutscher. Handwritten correspondences between male and female representatives of Ernestine Saxony in the early modern period (1550-1750) were digitized, catalogued, transcribed in full text, linguistically annotated on up to 17 different levels and analysed for phenomena like genderlect, language change, dialect, grammar, lexis, orality-literacy and signals of modesty and politeness. The annotated corpus has been published and provided for further linguistic research under open access conditions. Currently, the letters are also edited in full text in order to make them accessible for historical and literary research as well. This paper gives an overview over the corpus' contents, describes the successive work procedures and the annotation spectrum, and finally points out its subsequent scientific usability and accessibility.

¹ fasshauer@em.uni-frankfurt.de.

Corpus description

A thorough investigation of gender-related language characteristics requires a sufficient amount of personally handwritten letters by both male and female authors. Larger collections of Early New High German correspondences between men and women have survived only in the family archives of the ducal houses. Although the duchesses hardly ever took part in the political government, they nevertheless fulfilled important dynastic functions which included the writing of letters in order to maintain and cultivate the relations with the members of both their original and their husband's families. Writing letters in their own hand instead of having them written by office clerks indicated a special personal affection and respect (Nolte 2000). As documents of dynastic relevance, the duchesses' letters have been kept along with their husbands' political correspondence. Many of them can still be consulted in the State Archives – in our case those of Weimar, Dessau, Coburg, Gotha, Dresden, Altenburg and Munich.

The act of writing, however, was generally regarded as strenuous and exhausting. Very few princesses were diligent and willing writers, especially since the reigning prince usually functioned as the communicative centre of the court and therefore carried out the biggest part of the correspondence himself. Only in situations when their husbands were unable to correspond for one reason or another, the women were forced to write more extensively and frequently. In the century after the reformation, however, Ernestine princesses of three successive generations were either temporarily separated from their husbands or, as widows, had to advocate their cause mostly on their own. Elector John Frederick I the Magnanimous (1503-1554) and his eldest son John Frederick II (1529-1595) both were taken into imperial custody and partially or completely dispossessed of their territories. While Electress Sibylla (1512-1554) as well as her daughter-in-law Elisabeth (1540-1594) could still correspond and confer with their husbands, Dorothea Susanna (1544-1592) and Dorothea Maria (1574-1617) stayed behind alone with their children after both their husbands had died at an early age. Out of their precarious circumstances, they wrote supplication letters to the elector and the emperor, negotiated their children's dynastic, confessional and educational rights with their guardians and asked their male relatives for legal assistance and financial support. The necessity of treating new topics made them go beyond the

Period	Total amount of letters	Letters by female writers	Letters by male writers
1546-1575	249	135	114
1576-1600	71	33	38
1601-1625	190	103	87
1626-1650	0	0	0
1651-1675	1	1	0
1676-1700	44	42	2
1701-1725	5	5	0
1726-1756	40	38	2

Table 1: Temporal distribution of the letters in the corpus.

usual epistolary formula of the period and adopt an individual style and spelling habits. Thus, 300 women's letters were compiled and complemented by just as many letters from their male correspondents in order to investigate their distinctive gender-specific features by comparison with their male contemporaries' diction and grammar. As the following generations of Ernestine dukes suffered less tragic fates, their wives felt much less need to correspond; moreover, French gradually became the dominant language at German courts. The corpus therefore mainly consists of correspondences from the 16th and early 17th centuries.

Database and text transcriptions

Unlike the political correspondences of the reigning princes, the duchesses' letters have only sporadically been consulted for research, let alone scientifically edited. Thus, very few holding institutions offer single entry recordings of these letters; especially the State Archives mostly register them only as entire correspondence bundles. All letters relevant to the project were therefore recorded and made searchable in the *Fürstinnen-Briefdatenbank (FileMaker)*. This database contains detailed sets of metadata concerning both the letters and their writers. Apart from the conventional correspondence data – sender, recipient, place/date of dispatch, object location and classification number – the database also records the correspondents' biodata, their dynastic status and the princesses' families of origin. Due to unambiguous and hyphenated abbreviations, complete correspondences can be searched either individually or combined with other letters directed to the same recipient or stemming from the same author. The database also provides relation-oriented search options regarding the protagonists' gender, family relationship, dynastic function and marital state. A short abstract of each letter summarizes its contents and its textual modules; it also comprises all mentioned names and places in their correct and full form as well as the respective GND reference numbers. Moreover, the database contains diplomatic full-text transcriptions which were done manually from digital facsimiles of the handwritten texts.

Linguistic annotation

On the basis of the transcriptions, the texts were annotated linguistically on 17 different levels. The annotation was carried out by means of the XML-based open source software EXMARaLDA (Schmidt *et al.*). Under preservation of the original word order, the original Early New High German texts were translated into standardized New High German according to the ten-volume Duden as a reference dictionary (Duden 2000). The insertion of a modernized text-level guaranteed not only the comparability of the texts across the centuries, but also facilitated their semi-automatic lemmatization and part-of-speech-tagging by means of the TreeTagger (Schmid 1994-) according to the 'Stuttgart-Tübingen-Tagset' (STTS) (Schiller *et al.* 1999). The manually revised TreeTagger-output was complemented by a morphological analysis according to the same guidelines and served as a basis for the further annotation of the corpus.

	0	1	2	3	4	5	6	7	8	9
[tok]	Freuntliche	herczallerlieste	gemal	Jch	hab	deyn	schreiben	gancz	freuntlichen	fernommen
[norm]	Freundliche	herzallerliebste	Gemahl	ich	habe	Dein	Schreiben	ganz	freundlich	vernommen
[lemma]	freundlich	herzallerliebst	Gemahl	ich	haben	dein	Schreiben	ganz	freundlich	vernehmen
[pos]	ADJA	ADJA	NN	PPER	VAFIN	PPOSAT	NN	ADV	ADJD	VVPP
[morph]	Pos.Fem.Nom.Sg	Sup.Fem.Nom.Sg	Fem.Nom.Sg	1.Sg.*.Nom	1.Sg.Pres.Ind	Neut.Acc.Sg	Neut.Acc.Sg		Pos	
[grfunct]	voc			subj	prcompl	acc-o		modadv		prcompl
[s_grfunct]										
[clause-st]				decl						
[complex]	A_0			A_1						
[graph]	t_d	cz_z_0_b	g_G_0_h	J_i	0_e	d_D_ey_ei	s_S	cz_z	t_d_en_0	f_v
[phon]					Apo					
[quot]										
[lex_gr]			WS		E1, Temp				E2	
[mod_polite]	n2-fam			n1-pn		n2-pn			x	
[plural]	salut			recep						
[mean]			Gemahlin							
[comment]					Perf					
[formal]	16r, KV	KV	MOV, GK, KV	GK, VV		GK, KV	GK	KV	GK	KV
[mann]										

Figure 1: Annotation levels in EXMARaLDA's score editor.

Since STTS was developed for modern German, its applicability to Early New High German required its modification and extension by 10 more tags; in addition, over 150 tags have been developed for marking the syntactical, graphematic, phonological, lexicological, phraseological and stylistic peculiarities (Fasshauer *et al.* 2013a). The identification and description of all graphematic and phonological deviations of the original Early New High German from the normalized New High German text enables the investigation of language change, dialectal influences and gender-specific writing habits. In a thorough syntactical analysis, the type (e.g. declarative, interrogative or imperative), form (e.g. subjunctive or pronominal) and grammatical function (e.g. adverbial or attributive) of each clause as well as the sentence complexity were defined. Moreover, the topics dealt with in the single letters are tagged according to early modern letter rhetoric or *ars dictandi*.

Apart from formal categories like salutation, date or signature, textual statements like compliments, requests, inquiries or exhortations were also marked. The author's self-references and his or her way of addressing the correspondence partner are focussed especially because they contain signals of modesty and politeness allowing an investigation of the respective genderlect. This newly created annotation system consisting of c. 160 tags forms the Jena Tagset (JeTS).²

Corpus analysis

As the person- and letter-related metadata recorded in the database have been imported into EXMARaLDA, the Corpus Manager Coma is able to form subcorpora on the basis of various metacriteria. For instance, all letters by a widowed princess to her brothers containing more than one page and having been dispatched from a certain place of residence in one and the same year can be pre-selected and saved individually for further processing.

2 For a similar approach to adapt STTS to historical German see Dipper *et al.* 2013.

The analysis tool EXAKT enables detailed search queries both on the text levels and the annotation levels either for the whole corpus or for pre-defined sub-corpora. Again, search results can be filtered, saved for further processing and exported to other applications like MS WORD or EXCEL. By applying Regular Expressions (RegEx) as filter criteria, the text levels can be searched for any combination of characters. Search results can be displayed on different levels at the same time; irrelevant items can be unselected and removed from the list.

Apart from linguistic research, the corpus can also support historico-cultural studies and literary text analysis. For instance, on the modesty-politeness-level, all the authors' self-references as well as their references to the recipient or to third persons may give insights into their self-positioning and self-staging in a sociological respect; similarly, the detailed annotation of rhetorical phenomena like salutations, compliments, threats or curses on the phraseological level allows not only for stylistic analyses and literary genre studies but also for examinations of social, dynastic and political relationships between the persons involved.

Thanks to the standardization and lemmatization levels, the texts can be searched regardless of their original spelling in full text for certain correspondence topics. Subjects of interest are for instance the everyday life of the early modern high nobility with regard to their daily routine, their social activities, their handicraft, diet, diseases and medication, exchange of gifts, child care, gardening, historical persons, places and buildings or early protestant religious practice.

Edition and access

The database also forms the basis of a digital reading edition which is presently being developed in the 'Universal Multimedia Electronic Library' (UrMEL) (Fasshauer *et al.* 2015). UrMEL is provided by the Thuringian State and University Library (ThULB) and ensures sustainability, data maintenance and open access. Its contents are integrated into comprehensive web portals like Europeana or Kalliope. The edition provides a synoptic presentation of the digital facsimiles and the letter transcriptions which are encoded according to the TEI P5 guidelines. Shortly, the lexicological annotation levels (standardized text, lemmatization, POS, morphological analysis and meaning) will also be integrated.

The annotated text corpus has been published in the LAUDATIO-Repository (Fasshauer *et al.* 2015) which provides long-term access and usage of deeply annotated information. It has also been integrated into the ANNIS database (Krause *et al.* 2016) which is equipped with a web browser-based search and visualization architecture for complex multilayer corpora, allowing for complex corpus analyses by combined search queries on different levels. Both platforms have specialized in linguistic research and are universally accessible. A detailed description of the corpus structure (Fasshauer *et al.* 2013a) and a project documentation (Fasshauer *et al.* 2013b) have been published on the project homepage. The Jena Tagset (JeTS), which was designed especially for the annotation of historical texts, has been provided for free reuse (*ibid.*). Moreover, a volume dedicated to the corpus analysis is under preparation and will shortly appear in print.

References

- Dipper, Stefanie, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. HiTS: Ein Tagset für historische Sprachstufen des Deutschen. In *JLCL* 28. 1. 1-53.
- Duden. *Das große Wörterbuch der Deutschen Sprache, 10 vols.* 2000. Mannheim: Bibliographisches Institut.
- Fasshauer, Vera, Rosemarie Lühr, Daniela Prutscher, and Henry Seidel (eds). 2014. *Fürstinnenkorrespondenz (version 1. 1), Universität Jena, DFG. LAUDATIO Repository*. Accessed March 3, 2017. <http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm>.
- Fasshauer, Vera, Rosemarie Lühr, Daniela Prutscher, and Henry Seidel (eds). 2013a. *Dokumentation der Annotationsrichtlinien für das Korpus 'Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum'*. Accessed March 3, 2017. http://dwee.eu/Rosemarie_Luehr/userfiles/downloads/Projekte/Dokumentation.pdf.
- Fasshauer, Vera, Rosemarie Lühr, Daniela Prutscher, and Henry Seidel (eds). 2013b. *Korpus-Aufbau der 'Frühneuzeitlichen Fürstinnenkorrespondenzen im mitteldeutschen Raum'*. Accessed March 3, 2017. http://dwee.eu/Rosemarie_Luehr/userfiles/downloads/Projekte/Korpusaufbau.pdf.
- Fasshauer, Vera, Henry Seidel, and Daniela Prutscher. 2015. *Digitale Fürstinnenkorrespondenzen. FSU Jena*. Accessed March 3, 2017. <http://archive.thulb.uni-jena.de/hisbest/servlets/solt/collections?q=%2Bcomponent%3A%22e48b1d7f-6b9f-48cb-be43-98122e9f9b21>.
- Krause, Thomas and Amir Zeldes. 2016. 'ANNIS3: A new architecture for generic corpus query and visualization.' In *Digital Scholarship in the Humanities* (31). <http://dsh.oxfordjournals.org/content/31/1/118>.
- Nolte, Cordula. 2000. 'Pey eytler finster in einem weichen pet geschieben. Eigenhändige Briefe in der Familienkorrespondenz des Markgrafen von Brandenburg (1470-1530).' In *Adelige Welt und familiäre Beziehung. Aspekte der 'privaten Welt' des Adels in böhmischen, polnischen und deutschen Beispielen vom 14. bis zum 16. Jahrhundert*, edited by Heinz-Dieter Heimann. Potsdam: Verlag für Berlin-Brandenburg. 177-202.
- Schiller, Anne, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS. IMS Stuttgart / Sfs Tübingen*. Accessed March 3, 2017. <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>.
- Schmid, Helmut. 1994-. *TreeTagger. Universität München*. Accessed March 3, 2017 <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.
- Schmidt, Thomas, Kai Wörner, Timm Lehmberg, and Hanna Hedeland. (n. d.) *EXMARaLDA. Werkzeuge für mündliche Korpora. CRC 538 'Multilingualism', University of Hamburg*. Accessed March 3, 2017. <http://www.EXMARaLDA.org/>.

Hybrid scholarly edition and the visualization of textual variants

Jiří Flaišman,¹ Michal Kosák²

© Jakub Říha³

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

At the Institute of Czech Literature of the ASCR, textual scholarship ranks among the traditional disciplines. Since 1998, research has been based in the Department of Textual Scholarship. The work of the Department is considered central to the development of the discipline in the Czech Republic for several reasons: (1) the editions its members produce (especially *Kritická hybridní edice* (Hybrid Scholarly Edition)), (2) its method-oriented research, (3) its work in the literary history field, (4) the development of 'Varianty' (Variants), a specialized series devoted exclusively to textual criticism, (5) regular seminars/colloquiums constituting the basis of scholarly discussion of the field of textual scholarship, and (6) cooperation/tuition at universities in Prague and Olomouc.

The involvement with the digital editing began in the 1990s. At that time, a new project fully exploiting the new media was born: the Czech Poetry Database (as the name suggests, it was inspired by the English Poetry Database). The CPD is a full-text database of 1,700 volumes of poetry books (written in Czech in the 19th century and in the beginning of the 20th century). Its core unit is the poetry book (usually the first edition), which is presented firstly in a literal (diplomatic) transcription and secondly as an edited and corrected text. Descriptions of a unit usually do not include any information about the unit's textual history or any comparison or collation with other readings.

1 flaisman@seznam.cz.

2 kosak@ucl.cas.cz.

3 riha@ucl.cas.cz.

This situation, together with the endeavor to pursue editorial work, fed the interest in textual variants and their presentation in the digital environment. Our approach to textual variants was shaped by pioneering works by Czech structuralists, such as Jan Mukařovský's paper 'Variant readings and stylistics' (1930). While studying the variability of texts, we strive to identify the textual differences, or more precisely whole planes (strata) of textual differences, which point to the formation or transformation of authorial poetics. If we use the terminology of Russian textual scholarship (represented by Boris Tomashevsky, Konstantin Barsht and Pjotr Torop), we can say that we seek to track both changes in particular textual features and changes in the structure of the text as a whole. To put it simply, we are interested in the processuality of the text, and we attempt to capture the vector or vectors of its changes. Our interest in variants is not primarily linguistic – although linguistic analysis lies at the basis of our work – but aesthetic: we seek to capture the dynamic or fluid nature of literary texts and to be able to visualize them adequately.

The effort to conceive each variant as an independent and individual work and therefore not to declare one of the variants 'canonical' fundamentally shaped the conception of the *Hybrid Scholarly Edition*. To disrupt the identity of the 'canonical' text, the *Hybrid Scholarly Edition* combines a book edition for general readers and a digital edition for academic readers.

The digital edition brings together and arranges or organizes all the textual variants, such as drafts, manuscripts, fair copies, all the types of prints (magazine, book) and corrected prints. Every text is presented to the reader in different ways: (1) as a facsimile, (2) as a transcription (in the case of manuscripts), (3) as a literal (diplomatic) edition of printed texts and (4) as a corrected and commented edition. Textual changes furthermore are registered in the apparatus section in the form of a synoptical reading.

The digital part of the *Hybrid Scholarly Edition* does not aim to establish authoritative canonical reading but rather aims to grasp the substantial fluidity of the text. This aspect of the digital edition is reinforced by the different editorial approach to the book edition. The book edition allows for a progressive editorial approach, consisting of corrections, orthographic alterations, etc., while the digital edition holds to a conservative editorial approach, consisting of identification and correction of obvious writing or printing errors or of a diplomatic transcription.

This approach is preserved even in the edition of non-textual materials. For example, the digital edition of Gellner's collected works also includes the author's art. Its presentation also reflects variability and offers not only finalized paintings and printed illustrations but also drafts and proofs. The digital edition also comprises the complete, full-text database of critical evaluation and secondary sources, related to edited texts.

The main emphasis of the *Hybrid Scholarly Edition* is on the presentation of variants. Important progress in this area was made in the third volume of the edition: Petr Bezuč's *Silesian Songs* (2015). It introduced a new module for the analysis of variants that allows the reader to compare chronological subsequent

variants and also to identify the differences between them and to measure so-called Levenshtein distance. These tools allow us to introduce new statistical methods into the research of the variants.

We seek to approach the variability of the texts from various directions and to represent it in different ways. For example, a draft with several layers is reproduced as a photographic image. It enables the reader to follow the spatial dimension of the manuscript or the dynamics of writing. Then the same text is presented as a transcription, therefore an interpretation of the relation between changes. On top of that, two chronologically subsequent variants can be visualized in a special interface, with highlighted mutual differences (this solution resembles Juxta Commons, which was not familiar to us at the time the edition was being made). The synoptical reading then interrelates all the intratext variants with other preserved textual sources.

By juxtaposing various editorial approaches – facsimile, transcription, diplomatic edition or corrected edition supplemented with a commentary – we do not just want to offer verified textual material for other scholars; we also want to create a set of differently interrelated and hierarchically arranged editions that takes advantage of the differently based approaches to text.

Finally two areas of textual variability will be stressed which still remain a challenge for us. The first set of problems is associated with the so-called ‘genetically last layer of the text’. Each manuscript is captured as a static facsimile (image) and then its intratextual variability is interpreted in a synoptical reading. Furthermore, we present texts in a separate window as a genetically final form. We intended to represent variants simply, without their intratextual history, in the way Sergej Bondi once prepared the works of Alexander Pushkin – interpreting the surviving sketches so that he always could offer a definitive version of the text to the reader. However, our initial decision has been challenged lately by the very fact that not every draft seeks a definitive form; some of them diverge in several, equivalent directions. In these cases the practical editorial work illustrates the dilemma between the strictly applied genetic approach and approaches which take into account authorial intention, which is not always clear.

The second set of problems is related to the isolation of texts in our edition. As already mentioned, the core unit is a single text, be it a poem, a short story or a journal article, so it is complicated for us to address the variability of the composition within larger units. In the last volume we tried to capture the variability of the composition of a collection which was published during the author’s lifetime more than 30 times; however, we failed to demonstrate transformations in juvenile manuscript collections of Karel Hlaváček. This is not the most significant result of the isolation of individual texts in our edition. To pursue our goal – description of the formation or transformation of authorial poetics – we need to be able to show the interdependence of changes between poems and to define a network of synchronous changes and their motivations. Here, at the core of our effort, the ability to represent textual variability seems insufficient and it is necessarily left to the commentary.

It is obvious that the *Critical Hybrid Edition* actually combines different editorial approaches, that it hides many different editions in the disguise of a single one. Our position stems from skepticism toward single editorial approach, from the knowledge of its inner limitations. We try to ensure that each approach is aware of its own implications, that it realizes its own methodological shadow, and also that it see where its strengths lie.

Burckhardtsource.org: where scholarly edition and semantic digital library meet

Costanza Giannaccini¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Burckhardtsource.org is a semantic digital library created within the project 'The European correspondence to Jacob Burckhardt', funded by the European Research Council and coordinated by Prof. Maurizio Ghelardi (Scuola Normale Superiore, Pisa). It concentrates on a timespan ranging from 1842 to 1897 and witnesses a period of significant cultural transformations. The platform has a functional and intuitive structure: thanks to a broad view on the entire collection, the research team has identified specific research areas of particular relevance and therefore has planned preferential access paths (*Collections* and *Highlights*) to offer alternative interpretations. Contents may be queried and displayed in several ways, thanks to different input modalities in data and information.

The platform uses specific tools for content semantic annotation: *Pundit* and *Korbo*. In addition to thematic anthologies and to Highlights, users can lead autonomous researches alternatively entering a search term, or using filters (operating at metadata and semantic annotations level). Filters work both independently, and in combination. Search results offer several visualisation options, while single letters are presented in two versions: a semantic and a philological one.

¹ costanza.giannaccini@sns.it.

Introduction

The concept lying behind the project ‘The European correspondence to Jacob Burckhardt’² complies with both the desire to overcome the limits imposed by a conservative archival structure, and to meet scholars’ and researchers’ needs in the investigation of this rich collection and of all the historical and cultural information here collected. More than half of the entire correspondence is currently available on the platform: 700 letters can be analysed and read in their double version. Soon the remaining letters will be added to the collection. The correspondence counts around 400 authors and more than 1100 letters sent to the Swiss art historian in a time span of 55 years, in a period of particularly fervent European history and culture.

The platform has a functional and intuitive structure: thanks to the overview on the entire collection, the research team has identified specific research areas with particular relevance and has therefore planned preferential access paths (*Collections* and *Highlights*) in order to offer alternative interpretations. These entry points suggest innovative interpretations following a thematic common thread, identifying six recurring themes in the entire correspondence. These topics

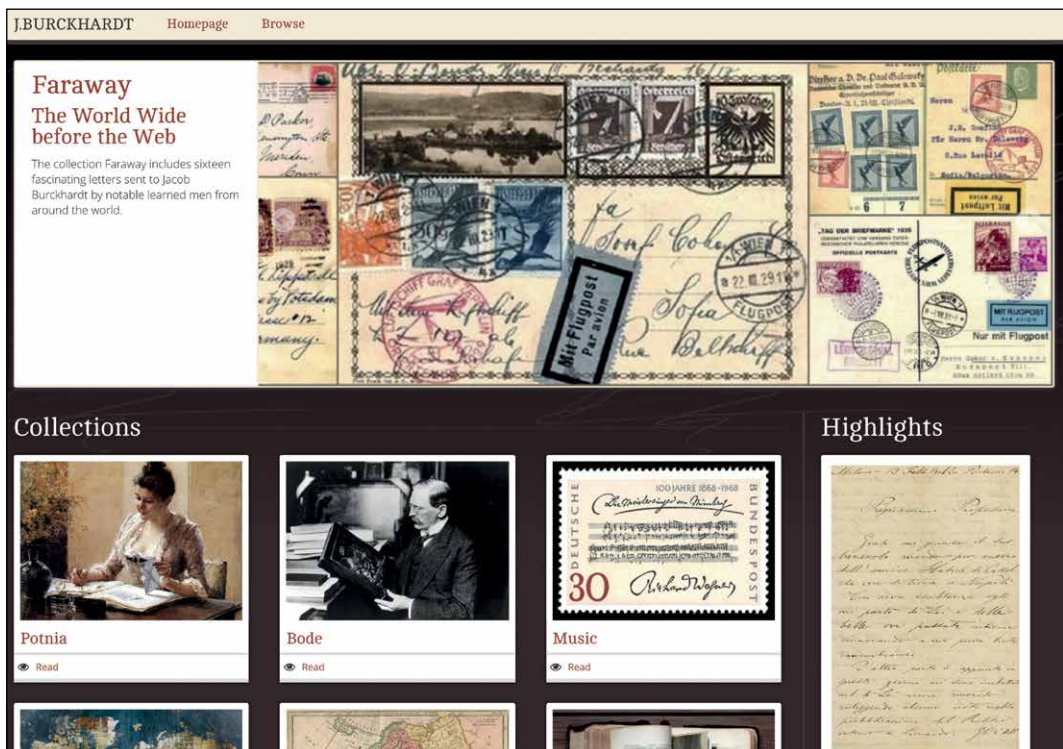


Figure 1: Burckhardtsource has a functional and intuitive structure. Collections and Highlights are preferential access paths offering alternative interpretations on the entire collection.

2 EUROCORR ERC Advanced Grant Project – Advanced Grant EUROCORR, Grant Agreement no. 249483.

flow into anthologies, bringing factual information on relevant historical and cultural issues. Besides, the insight into the entire collection suggests a further access point on an inter-thematic level. *Highlights* propose a selection of letters with notably historical and cultural importance, showing significant facts related to both territory and society (Figure 1).

Contents are queried and displayed in different ways, thanks to the input modalities of data and information. Information on letter materiality is detailed in the metadata section, while semantic meaning is added through annotation. Data belonging to metadata and annotations are easily retrievable with the search tool.

Annotations are integrated via the web annotator tool *Pundit* in combination with the internal vocabulary and semantic web basket manager *Korbo*. Queries have two alternative directions: one search term may be added in the *Browse* space, otherwise special filters lead information retrieval by means of appropriate selections.

Visualisation has a fundamental role. Fully conscious of the importance of both data retrieval and visualisation potentialities, Burckhardtsource.org offers different systems for contents display. Search results – more or less numerous depending on the filters set – are visible spatially on a map, or chronologically through a timeline. A final, and particularly innovative visualisation modality is the *Advanced Semantic Search*. The link on the *Homepage* is a random entry point, showing at each new access a different sender: this semantic search displays further connections with other people, places and art works, t.i. with other semantic contents.

Entry points

Burckhardtsource.org homepage is divided into different areas corresponding likewise to access points. The main part is reserved to *Collections*, i.e. anthologies collecting letters tied to specific themes. One of those, *Potnia* analyses women's social condition in the second half of the 19th century: themes emerge through the pen of Lady Writers and male correspondents. *Music* gathers letters from correspondents sharing with Burckhardt the same passion for music. The *Faraway* collection shows how small the world used to be in the nineteenth century, collecting letters sent to the Swiss art historian by notable learned men from around the globe. Important evidence on the major European events during the second half of the century is highlighted in the *Europe Collection*. In *Photography* Burckhardt's response and use of the new medium proves how central this topic was in the period. Finally, in comparison to all the other collections, *Bode* is unique, since it stores the complete known correspondence between Jacob Burckhardt and Wilhelm von Bode.

The selection of these thematic areas has been possible thanks to a broad view on the entire collection of letters. This same perspective led to the creation of another 'super-collection', the *Highlights*, another access point to the platform. This selection does not follow a thematic criterion: the letters collected here have value in themselves and excel over the whole correspondence.

Tools

Pundit and Korbo

The platform uses specific tools for content semantic annotation. *Pundit* is an annotator working in association with the web basket manager *Korbo*, a customised dictionary containing annotated entities present in the letters. Once semantic value is added, content is retrieved through four annotated categories: People, Places, Artworks and Bibliographic citations.

Search

The search tool *Browse* is recommended for free searches in both metadata and semantic annotations. Moreover, the platform added value in research is the filters tool: on the left side of the screen are several research categories – Year, Sender, Compilation and Receiving place – coming from the metadata, while on the right side are aligned the previously mentioned semantic categories. Users therefore can lead autonomous researches through filter selection, by selecting specific information coming from the metadata (e.g. Year), from semantic annotations (e.g. Artwork), or combining more filters (e.g. selecting a Compilation place and a Bibliographic indication) (Figure 2).

The screenshot displays the Pundit/Korbo search interface. On the left, the 'Filters' section includes metadata categories: Years (1893: 2), Senders (Lendorff, Hans: 2), Compilation place (Anticoli Corrado: 2), and Receiving place (Basel: 2). The top search bar shows 'Search results: 2' and a search input field. The main content area displays two search results, each with a thumbnail of a handwritten letter and a detailed description. The first result is 'From: Lendorff, Hans To: Burckhardt, Jacob, Anticoli Corrado, 1893, letter 782', and the second is 'From: Lendorff, Hans To: Burckhardt, Jacob, Anticoli Corrado, 1893, letter 802'. On the right, the 'Speaks of' section lists semantic categories: Persons (Ingres, Jean-Auguste-Dominique: 1, Lendorff-Berri, Hanna: 1, Preiswerk, Carl Andreas: 1, Raffaello, Sanzio: 1, Rembrandt, Harmensz van Rijn: 1), Places (Anticoli Corrado: 2, Basel: 2, Rom: 2, Sabiner Berge: 2, Aarburg: 1), Artworks (Caffè Greco, Rom: 1, Cappella Sistina (Rome): 1, Michelangelo Buonarroti, Deckenbild, Cappella Sistina: 1, Piazza di Spagna, Rom: 1, Raffaello, Parnace: 1), and Bibliographies (Burckhardt, J., & Burckhardt, M. (1949-1994), Briefe: 2).

Figure 2: On the left side of the screen are metadata research categories, on the right side are aligned four semantic categories: People, Places, Artworks and Bibliographic citations. Users can lead autonomous researches through single or combined filter selection.

Visualisation

In addition to the different research and selection types, Burckhardtsource.org offers several visualisation options of the results. Indeed, letters might be ordered chronologically on a timeline, or on a map in case of users interested in the geographic issue.

Eventually, an innovative visualisation system is the *Advanced Semantic Search*. The access point in the Homepage is a completely random threshold to the entire letters collection: it is at the meantime an additional access point, since it reveals further connections among letters. The dots around the access entity are likewise connections to other entities, therefore showing additional links, unknown to more traditional search typologies (Figure 3).

Each letter on the platform has two versions: a semantic and a philological one. The first version is a running text showing semantic annotations, specifically conceived for users more inclined to letters content. The second is a scholarly edition containing all the text variants that users more interested in the philological aspect of the correspondence can easily and autonomously select in the drop down menu.

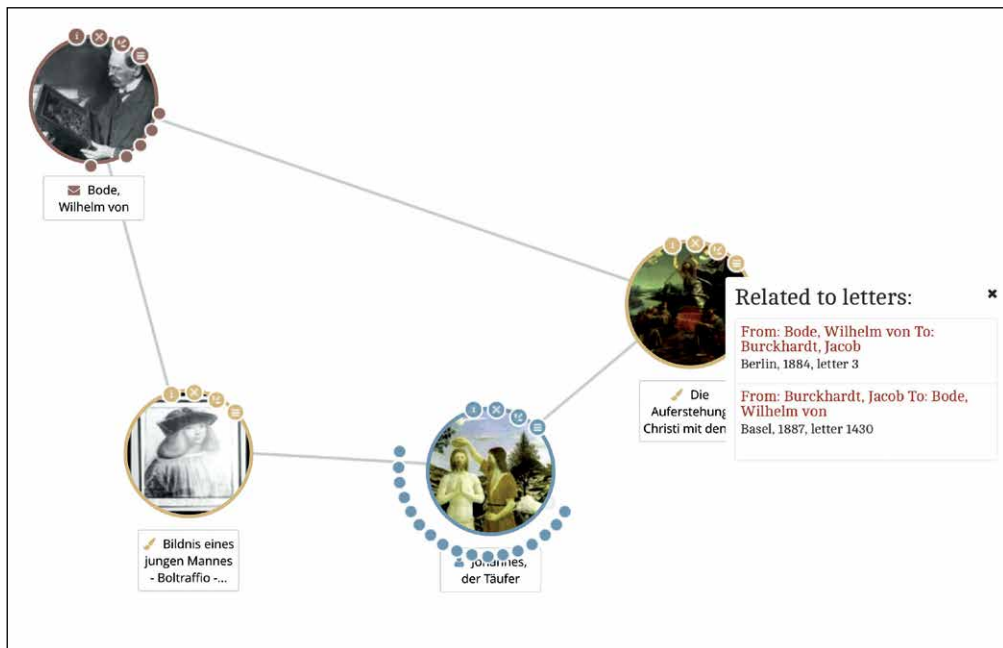


Figure 3: The *Advanced Semantic Search* is a completely random threshold to the entire letters collection. Compared to more traditional tools, this search shows additional links among letters.

Conclusions

In conclusion Burckhardtsource.org is an advanced and enhanced scholarly edition where data and information input can be retrieved, queried and visualised independently, in accordance with different research needs. The digital tools suggest alternative perspectives on the entire collection, opening new research panoramas. The critical rigour in texts analysis, in metadata collection and in annotations is certified by the copious documentation available on the site.

References

- Blog des Staatsarchivs Basel-Stadt*. Accessed March 3, 2017. <http://blog.staatsarchiv-bs.ch>.
- Burckhardtsource. Accessed March 3, 2017. <http://burckhardtsource.org/>.
- Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg*. Accessed March 3, 2017. http://www.fotomarburg.de/forschung/laufende_projekte/burckhardt?searchterm=burckhardt.
- Deutsche Nationale Bibliothek: Standardisierung*. Accessed March 3, 2017. http://www.dnb.de/DE/Standardisierung/standardisierung_node.html;jsessionid=A5ECE-056DAE8D3425548CBBD12119.prod-worker3#doc1428bodyText3.
- Di Donato, Francesca and Susanne Müller. 2014. 'Burckhardtsource.org A semantic digital edition of the correspondence to Jacob Burckhardt'. *Lexicon Philosophicum* 2: 327-335.
- Di Donato, Francesca. 2013a. 'Semantic annotation of Digital Libraries: a model of science communication'. *Presentation given at the University of Pisa 23-24 May 2013*. Accessed March 7, 2017. <https://www.slideshare.net/FrancescaDiDonato/semantic-annotation-of-digital-libraries-a-model-for-science-communication>.
- . 2013b. 'Working in Scholarly Content: a Semantic Vision'. *Paper presented at 'Open Platforms for Digital Humanities'*, Cortona, Scuola Normale Superiore, 17 gennaio.
- Fonda, Simone. 2014. 'Pundit, an Open Source semantic annotation tool for the web'. *Presentation at Centre Alexandre Koyré Histoire des Sciences et des Techniques, Paris*. Accessed March 3, 2017. <http://www.slideshare.net/netseven/pundit-an-open-source-semantic-annotation-tool-for-the-web>.
- Giannaccini, Costanza. 2014a. 'Burckhardtsource.org – The European correspondence to Jacob Burckhardt'. *Paper presented at Datenmodellierung in digitalen Briefeditionen und ihre interpretatorische Leistung: Ontologien, Textgenetik und Visualisierungsstrategien, Berlin, 15-16 May*.
- . 2014b. 'Burckhardtsource.org: a Semantic Digital Library'. *Paper presented at AIUCD: La metodologia della ricerca umanistica nell'ecosistema digitale, Bologna, 18-19 September*.
- . 2014c. 'Enriching the Web of Data with Artworks: Burckhardtsource.org experience'. In *Electronic Imaging & the Visual Arts (EVA), Conference Proceedings, Florence, 7-8 May*, edited by Vito Cappellini, Florence, Florence University Press. Accessed March 3, 2017. http://www.fupress.com/archivio/pdf/2759_6474.pdf.

- Korbo. *The semantic basket manager*. Accessed March 3, 2017. <http://www.korbo.org>.
- Müller, Susanne, and Francesca Di Donato. 2013. 'Burckhardtsource.org A semantic digital edition of the correspondence to Jacob Burckhardt'. In *EVA-Berlin Conference Proceedings*. Accessed March 3, 2017. https://www.academia.edu/9940791/Burckhardtsource.org._A_semantic_digital_edition_of_the_correspondence_to_Jacob_Burckhardt.
- Müller, Susanne. 2013. 'Burckhardtsource: The European Correspondence to Jacob Burckhardt'. *Paper presented at Workshop: Open platforms for digital humanities, Cortona, 26-27 September*.
- Online-Edition Briefe an Jacob Burckhardt*. Accessed March 3, 2017. <http://www.staatsarchiv.bs.ch/news/2016-04-18-online-edition-burckhardtsource.html>.
- Pundit. Accessed March 3, 2017. <http://www.thepund.it>.

EVI-linhd, a virtual research environment for digital scholarly editing

*Elena González-Blanco,¹ Gimena del Río,
Juan José Escribano, Clara I. Martínez Cantón
& Álvaro del Olmo*

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Digital Humanities, as a scientific field, can be seen as a boundary discipline that requires cooperation and common agreements and views among many scientific communities (del Río Riande 2016). There are some tools that facilitate communication and understandings across different areas and even projects. These are what in sociology have been called boundary objects, described by Star and Griesemer (1989, 393) in this way:

Boundary objects are objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use.

This concept is crucial when talking about collaborative and interdisciplinary labour. Virtual Research Environments (VREs) have become central boundary objects for digital humanists community, as they help global, interdisciplinary and networked research taking of profit of the changes in 'data production, curation and (re-)use, by new scientific methods, by changes in technology supply' (Voss and Procter 2009, 174-190). DH Centers, labs or less formal structures such as

¹ egonzalezblanco@flog.uned.es.

associations benefit from many kind of VREs, as they facilitate researchers and users a place to develop, store, share and preserve their work, making it more visible. The implementation of each of these VREs is different, as Carusi and Reimer (2010) have stated, but there are some common guidelines and standards generally shared.²

This paper presents the structure and design of the VRE of LINHD, the Digital Innovation Lab at UNED³ and the first Digital Humanities Center in Spain. It focuses on the possibilities of a collaborative environment focused on a very realistic type of research: a non-English speaker, relatively new in DH technologies, which is keen on working in his project with his team, but does not have a uniform team of researchers (that means, they have different levels of understanding DH technologies).

Taking into account the language barrier that English may suppose for a Spanish-speaking scholar or student and the distance they may encounter with the data and organization of the interface (in terms of computational knowledge) while facing a scholarly digital edition or collection, LINHD's VRE comes as a solution for the virtual research community interested in scholarly digital work.

The main aims of EVI are:

- Promoting digital scholarly editions in Spain, as well as the humanist training in the field of Digital Humanities through the use of standards (such as TEI-XML), distinguishing the three fundamental processes involved in the development a complete digital edition: text tagging, analysis, text processing, and finally visualization and digital publication.
- Managing through digital tools and databases text collections that contain tagged texts (displaying different visualization possibilities) and link with other non-text content (such as images or multimedia content) labeled with metadata.
- Enabling recovery of such content.
- Providing the humanist researcher the building of digital repositories in the cloud using technologies of the semantic web and linked data (LOD) allowing standardization of content and interoperability with other projects, resources and databases.

In this sense, our project dialogues and aims to join the landscape of other VREs devoted to digital edition, such as *Textgrid*, *e-laborate*, *ourSpaces*, etc. and, in a further stage, to build a complete virtual environment to collect and classify data, tools and projects, work and publish them and share the results. Therefore, the key of our VRE is the combination of different open-source software that will enable users to complete the whole process of developing a digital editorial project. The environment is, up-to-now, divided into three parts:

2 As an example, see the Centernet map (<https://dhcenternet.org/centers>) and guidelines of TGIR Huma-Num 2015 (<http://www.huma-num.fr/ressources/guides>).

3 <http://linhd.uned.es>.

1. A repository of data to (projects, tools, etc.) with permanent identifiers in which the information will be indexed through a semantic structured ontology of metadata and controlled vocabularies (inspired in LINDAT, Isidore and Huni).⁴
2. A working space based on the possibilities of eXistDB to work on text encoding, storing and querying, plus some publishing tools (pre-defined stylesheets and some other open-source projects, such as Sade, Versioning machine, etc.).
3. A collaborative cloud workspace which integrates a wiki, a file archiving system and a publishing space for each team.

The impact of EVI-LINHD resides in building a very useful tool for the development of the humanities studies within a digital society. It aims to facilitate the change of the traditional editor's job to a virtual environment where accessibility, dissemination and visualization possibilities of the cultural object greatly increase the prospects of their study. A platform of this kind, pioneer in the Spanish-speaking community, will also facilitate the interoperability of our projects in international groups and networks working on similar topics.

EVILINHD is a powerful cloud-based platform that will offer researchers a space to manage their projects from the beginning to their publication and dissemination period, all through a single interface, which is thought of and designed as the key for the success of such a project: the research user.

⁴ The ourSpaces Virtual Research Environment project have worked in this sense developing an extensible ontological framework for capturing the provenance of the research process that they describe in Edwards (et al. 2014).

References

- Boot, Peter. 2012. 'Some digital editions and some remaining challenges.' *JANUS* 1: 39-54. Accessed November 29, 2015. <http://hdl.handle.net/2183/12646>.
- Candela, Leonardo. 2011. 'Virtual Research Environments.' *Scientific report. GRDI2020*. November 2. Accessed October 28, 2015. <http://www.grdi2020.eu/Repository/FileScaricati/eb0e8fea-c496-45b7-a0c5-831b90fe0045.pdf>.
- Carusi, Annamaria, and Torsten Reimer. 2010. 'Virtual Research Environment Collaborative Landscape Study. A JISC funded project'. Report, January. Accessed October 28, 2015. <https://www.jisc.ac.uk/rd/projects/virtual-research-environments>.
- Del Río Riande, Gimena. 2016. 'Humanidades Digitales. Construcciones locales en contextos globales'. SEDICIBlog, March 22. Accessed March 28, 2016. <http://sedici.unlp.edu.ar/blog/2016/03/22/humanidades-digitales-construcciones-locales-en-contextos-globales/>.
- Edwards, Peter *et al.* 2014. 'Lessons learnt from the deployment of a semantic virtual research environment.' *Web Semantics: Science, Services and Agents on the World Wide Web* 27, 70-77. Accessed April 28, 2016. <http://www.sciencedirect.com/science/article/pii/S1570826814000560>.
- Llewellyn Smith, Christopher *et al.* 2011. *Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century*. London: The Royal Society.
- Schlitz, Stephanie A., and Bodine S. Garrick. 2009. 'The TEIViewer: Facilitating the transition from XML to web display.' *Literary and linguistic computing* 24 (3), 339-346.
- Star, Susan Leigh, and James R. Griesemer. 1989. 'Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39.' *Social studies of science* 19 (3), 387-420.
- TCIR Huma-Num. *Le guide des bonnes pratiques numériques* (version of 13. 1. 2015). Accessed October 28, 2015. <http://www.huma-num.fr/ressources/guide-des-bonnes-pratiques-numeriques>.
- Voss, Alexander, and Rob Procter, 2009. 'Virtual research environments in scholarly work and communications'. *Library Hi Tech*, 27 (2), 174-190.

Critical diplomatic editing

Applying text-critical principles as algorithms

Charles Li¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

In recent years, text-based research in the humanities has shifted dramatically from working with critically edited texts to diplomatically-transcribed documents. This is with good reason: both the refinement of computational techniques and the growing interest in the intricacies of textual transmission have led scholars to create archives of transcribed documents in order to facilitate computer-aided textual analysis. But for scholars of an ancient text, for which no autograph exists, it is still vitally important to have a critical edition, with a carefully-curated apparatus, to work with. This is especially evident in the context of Sanskrit texts, some of which exist in dozens, if not hundreds of manuscript witnesses, most of which are extremely corrupt. Many of these documents, when transcribed diplomatically, are simply unreadable.

But nothing prevents us from producing both a critical edition and an archive of document transcriptions that are the source of the edition; in fact, this seems like a natural solution, not only because it facilitates corpus research, but also because it makes the edition much more transparent and open. As a scholarly product, the critical edition should be, in a way, reproducible – the reader should be able to trace an edited passage back to its sources easily, noting precisely what emendations have been made. The critical apparatus traditionally has served as the repository for this information, but, crucially, some silent emendations and omissions – for example, of very common orthographic variants – inevitably are made in order to make the apparatus useful. However, the ‘usefulness’ of a critical apparatus depends both on the editor’s judgment of what to include or exclude and also on a given reader’s needs, which may or may not align with the editor’s

¹ cchli@cantab.net.

critical principles; for example, while the editor may be trying to reconstruct the text as it was composed by the author, the reader may be trying to understand the text as it was read by later commentators. The challenge, then, is to make the critical apparatus flexible – to allow the reader to change the level of detail presented in the apparatus, on demand. Machine collation, applied to diplomatic transcripts, can produce a completely unselective, uncritical apparatus; however, when the collation algorithm is parameterized with a set of critical principles, then a selective apparatus can be generated which can be refined by the reader according to their needs.

The *Dravyasamuddeśa* project

The *Dravyasamuddeśa* project currently is producing an online, digital edition of the *Dravyasamuddeśa* of Bhartṛhari, a Sanskrit text on the philosophy of language, along with the *Prakīrṇaprakāśa* commentary by Helārāja. In order to achieve the aim of realizing an ‘open source’ edition, each witness is transcribed diplomatically in TEI XML and linked to the edition text.² These witnesses then are collated automatically, using the *Myers diff* algorithm (Myers 1986, 251-266), to produce an apparatus. However, since the diplomatic transcripts contain variations in punctuation, orthography, and the application of sandhi rules, the *diff* algorithm naively would report these differences in the apparatus. Therefore, in order to refine the generated apparatus, the web interface of the edition includes a number of options to filter out unwanted information (Figure 1). By using a machine collation algorithm rather than collating manually, the results are more consistent and precise, since a great deal of human error is avoided. Moreover, an apparatus can be generated automatically for any witness as a base text; the critical text no longer has the same privileged status as in print editions, where the witness texts exist only as apparatus variants. All of the diplomatically transcribed witnesses are treated as texts in their own right and are fully searchable. But perhaps most importantly, working with diplomatic transcripts and machine collation forces the editor to express their text-critical principles in a precise and formal manner, as machine-readable algorithms.

Text-critical principles as regular expressions

In order to filter out unwanted entries from the automatically-generated apparatus, the diplomatic transcripts are pre-processed prior to being collated. The pre-processing is performed in three stages: first, XML tags are stripped, along with tagged content that should be ignored (such as marginal notes and deleted text); secondly, punctuation as well as other irrelevant characters, such as digits, are removed; and finally, the orthography is normalized according to a set of text-critical principles. This last operation, normalization, is achieved by expressing the text-critical principles as regular expressions.

2 See Formigatti (forthcoming), section 3. 2, for an overview of applying TEI to South Asian manuscript sources.

Generate apparatus

Other witnesses ▾

Options

XML tags ▾

Punctuation ▲

☒ ignore abbreviation sign 「◦」

☒ ignore avagrahas 「|」

☒ ignore brackets

☒ ignore commas

☒ ignore daṇḍas

☒ ignore empty śirorekha 「—」

☒ ignore explicit hiatus 「_」

☒ ignore hyphens and dashes

☒ ignore line fillers 「|」

☒ ignore middot 「·」

☒ ignore numbers

☒ ignore puṣpikā 「❧」

☒ ignore periods/ellipses

☒ ignore quotation marks

Orthographic variants ▾

Figure 1: Apparatus options.

Regular expressions are a way of formally describing a search pattern, and they are implemented in most programming languages.³ For the purposes of normalizing a text for machine collation, regular expressions can be used to replace orthographic variants with their normalized counterparts. In a typical print edition, the text-critical principles that dictate what kinds of variation are ignored are stated in the preface, and those principles are applied silently as the editor collates the witnesses. Even digital projects that use computer software to analyze textual variation usually emend the source texts, rather than work with diplomatic transcriptions; for example, take this recent project at the University of Vienna that aims to produce a critical edition of the *Carakasamhitā Vimānasthāna*:

³ See Chapter 9 of *The Open Group Base Specifications, Issue 7*.

In the first phase of our still-ongoing editorial work, the ‘collation,’ all textual witnesses are compared with the widely known edition of Trikamji, that we chose as our standard version. In the course of this comparison all differences in readings between the manuscripts and the text as edited by Trikamji are noted with very few exception, like, for example, sandhi-variants, variants of punctuation, variants of consonant gemination after ‘r,’ variants of homograph and semi-homograph akṣaras

(Maas 2013, 32).

Just as in a traditional critical edition, the witness texts are collated manually, and some variants are discarded completely. But if we employ machine collation, we can transcribe diplomatically all witness texts, and then use text-critical principles, precisely expressed as regular expressions, to normalize them before collating.

Example: normalizing semi-homograph nasals

One common variation that is ignored in critical apparatuses of Sanskrit texts is that of semi-homograph nasals. In most scripts used to write Sanskrit, the nasals *ñ*, *ṇ*, *ṅ*, and *ṇ*, along with *m*, often are written as the anusvāra *ṁ*. In most editions, this variation is discarded silently; typically, the editor would express this rule in a prefatory statement, such as ‘variants of semi-homograph nasals are not noted’. But with machine collation, this rule must be expressed in a formal language, and this requirement gives us the opportunity to refine our text-critical principle to be as specific as possible. The replacement of a nasal with *ṁ* occurs only under certain specific conditions, and, based both on Sanskrit grammatical theory and a survey of the manuscripts being collated, a formal rule can be devised which expresses these conditions. In the case of semi-homograph nasals, the text can be normalized using the regular expression

$$/\dot{n}(?=[kg])|\dot{n}(?=[cj])|\dot{n}(?=[\text{ṭḍ}])|n(?=[\text{tdn}])|m(?=[pb])/ṁ/$$

Expressed in English, this means:

Replace

ñ when followed by *k* or *g*,

ñ when followed by *c* or *j*,

ṅ when followed by *ṭ* or *ḍ*,

n when followed by *t*, *d*, or *n*,

and *m* when followed by *p* or *b*,

with *ṁ*.

When this regular expression is applied to the texts before they are compared by the *diff* algorithm, the resulting apparatus will not include semi-homograph nasal variants. This approach is preferable to manual collation in a number of respects: firstly, expressing a text-critical principle in a formal language such as a regular expression forces the editor to be as specific and precise as possible; secondly, the diplomatic transcript of the manuscript, with its own particular orthography, is unaffected by the process; and finally, any one of these rules can be turned

off by the reader, resulting in, for example, semi-homograph nasal variants being included in the automatically generated apparatus. As a result, the apparatus that is produced is both precise and flexible.

Conclusion

In 1973, Martin L. West declared machine collation not worthwhile, criticizing it for producing ‘a clumsy and unselective apparatus’ (West 1973, 71-72). His criticism is strictly correct, and, in fact, it articulates a general problem in data-driven analysis: datasets, even very large ones, contain inherent biases, and a straightforward analysis would simply reproduce those biases.⁴ In order to achieve meaningful results, domain-specific knowledge needs to be applied. In the example of normalizing semi-homograph nasals, domain-specific knowledge was acquired – gleaned from Sanskrit grammar as well as experience working with Sanskrit manuscripts – expressed formally as a regular expression, and used as a pre-processing step to a general-purpose algorithm, *Myers diff*. By applying text-critical principles to the task of machine collation, an apparatus can be generated automatically that is neither clumsy nor unselective, and which is more precise than what could have been achieved manually.

Since West made his statement criticizing machine collation, there has been a shift in scholarly attitudes towards what a critical edition is and what it means for an apparatus to be ‘selective’. As West himself admits, editors cannot always be trusted, and the critical apparatus is a way for the reader to check the assertions of the editor. But the apparatus itself is also curated by the editor, and it serves to restrict the reader to a very limited perspective of the textual evidence for the edition. For the scholar of an ancient text, this is not enough; new modes of inquiry demand access to more and more information about the source material. In the *Dravyasamuddēśa* project, we hope to facilitate this by making the edition as ‘open source’ as possible, without sacrificing the intelligibility of a ‘selective’ critical apparatus; we merely have expressed our selection criteria – our text-critical principles – as filters that can be turned on or off by the reader. In effect, the apparatus is transformed from a static, authoritative presentation of textual evidence to the site of a negotiation between the textual hypothesis of the editor and the analysis of the reader.

सर्वभाषेषु ब्रह्मणो द्रव्यलक्षणस्याभेदात्तदभिधायित्वे शब्दानां सर्वत्र तस्य भावात् **सार्वार्थ्यं** शब्दान्तराभिधीयमानार्थत्वं साङ्ख्यं प्रसज्येतेत्यत्रेदमुच्यते । प्रतिनियताकारपरिच्छिन्नवृत्तित्वात्सर्वार्थत्वप्रतिबन्धादसङ्कर इत्यर्थः ।

G₁, G₂: सर्वतावेष्टु H: °णा G₁, G₂: °द्वत् G₁, G₂: तदतिधा-
तित्वे G₁, G₂: भावा P: स° [L: शब्दा° G₁, G₂: श-
ब्दान्तराति° K, V: शब्दान्तराभिधा°] K, V: सार्क्यं
[G₁, G₂: प्रस° V: प्रसज्येतेत्य° A: प्रसज्यतेत्य° P: प्रसज्यत
त्य° H: प्रसज्येतेत्य°] G₁, G₂: अत्रेदन् [L: प्रतिनियत°
G₁, G₂: °छिन्नवृत्तित्वा°] [V, P: सर्वार्थत्वं° C₇: सर्वार्थ°
K°: सर्वार्थत्वा°]

Figure 2: The generated apparatus: clicking on the variant highlights the lemma in the text.

⁴ For examples from research in the humanities, see Gitelman 2013.

Resources

The software being developed is based on open source libraries and is itself open source; the code is hosted on GitHub:<<https://github.com/chchch/upama>>. An online demonstration of the edition-in-progress can be found at <<http://saktumiva.org/wiki:dravyasamuddesa:start>>.

References

- Formigatti, Camillo A. Forthcoming. 'From the Shelves to the Web: Cataloging Sanskrit Manuscripts in the Digital Era'. In *Paper & Pixel: Digital Humanities in Indology*. Edited by Elena Mucciarelli and Heike Oberlin. Wiesbaden: Harrassowitz.
- Gitelman, Lisa, (ed.) 2013 *'Raw Data' is an Oxymoron*. Cambridge, Mass.: MIT Press.
- Institute of Electrical and Electronic Engineers and The Open Group. 2016 'Regular Expressions'. *The Open Group Base Specifications*, Issue 7. Accessed 4 March 2017. http://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap09.html.
- Maas, Philipp A. 2013. 'On What to Do with a Stemma – Towards a Critical Edition of the Carakasamhitā Vimānasthāna 8. 29'. In *Medical Texts and Manuscripts in Indian Cultural History*, edited by D. Wujastyk, A. Cerulli and K. Preisendanz. New Delhi: Manohar.
- Myers, Eugene W. 1986. 'An O(ND) Difference Algorithm and its Variations.' *Algorithmica* 1: 251-266.

St-G and DIN 16518, or: requirements on type classification in the Stefan George edition

Frederike Neuber¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

This paper reports on research emerged during the creation of a digital scholarly edition of Stefan George's (1868-1933) poetical works. Since 1904, George published his literary works in a particular typeface: the *Stefan-George-typeface* (St-G). The type design of St-G broke with typographical conventions at the time by including formal and stylistic features of historical and foreign scripts. Since this reference to past times and cultures is related closely to George's poetical motifs, typographical features need to be analyzed in coherency with the history of typography. Consequently, a classification of the St-G-types needs to be included in the digital edition. Since scholarly editing practices do not offer a standard or best-practice model for type classification yet, this paper takes a closer look at a type classification from the field of graphic design and typography: the German standard type classification 'DIN 16518'. After giving a concise overview of type classification in general, the paper will examine the DIN standard and evaluate it by attempting to classify St-G-types. In conclusion, against the background of the conference title 'Technology, Software, *Standards*', the paper will discuss requirements for the creation of a sustainable standard of type classification, which allows for the identification and classification of the St-G-typeface.

¹ neuber@i-d-e.de.

Type classification

Classification systems are powerful forms of knowledge organization. A classification is ‘the assignment of some-thing to a class; (...) it is the grouping together of objects into classes. A class, in turn, is a collection (...) of objects which share some property’ (Sperberg-McQueen 2004: ch. 14 §1). Since the purpose of a classification determines the choice of the distinctive properties that define the classes, it always has to be perceived in relation to when, by whom and for what purpose it was created.

The first serious endeavor of a type classification was made by the Frenchman Francis Thibaudeau in 1921. He grouped types based on their serifs, which up to this stage in typography were the most distinctive feature (Kupferschmid 2012). In 1954 – because of the increasing variety of types – the Frenchman Maximilian Vox proposed an expanded version of Thibaudeau’s system, which would become the cornerstone for future standard type classifications: ‘Vox/ATypI’ by the *Association Typographique Internationale* (1962), ‘Typeface Classification 2961’ by the *British Standards* (1967) and ‘DIN 16518’ (hereinafter called ‘DIN standard’) by the *German Institute for Standardization* (1964).

The DIN 16518 standard

The purpose of the DIN standard and its national counterparts is to sort types according to their formal properties with regard to their historical development. The type classes are built according to a number of distinctive properties of the type’s anatomy: terminals (serifs or sans serif), form of the serifs, thickness of the strokes, symmetry of the curve’s axis, cross stroke of the lower case <e>, form of the leg of <R>, tail of <g>, and morphology of <a>. The DIN norm divides types into eleven classes and five subclasses (see Figure 1): classes I-IV are named after the periods when the types were designed and they are distinct through the characteristics of the serifs. Group V also contains serif-typefaces but in this case, the rectangular slab serifs. Class VI ‘Serifenlose Linear-Antiqua’ (for English translations see Figure 1) includes types without serifs but with linear strokes. Class VII ‘Antiqua-Varianten’ combines all sans-serif variants which do not belong in any other class. Class VIII ‘Schreibschriften’ refers to typefaces which evoke the formal penmanship or cursive writing while class IX ‘Handschriftliche Antiqua’ contains handwritten – but not connected – types based on Antiqua forms. The tenth group ‘Gebrochene Schriften’ is a specific national class containing blackletter scripts. It is the only class containing subclasses. The eleventh and last class ‘Fremde Schriften’ includes writing systems which are not based on the Latin alphabet such as Arabic, Cyrillic and Greek (paragraph summarized from Runk 2007: 46-58).

I	Venezianische Renaissance-Antiqua (Humanist)
II	Französische Renaissance-Antiqua (Garalde)
III	Barock-Antiqua (Transitionals)
IV	Klassizistische Antiqua (Didone)
V	Serifenbetonte Linear-Antiqua (Slab Serif)
VI	Serifenlose Linear-Antiqua (Sans Serif, Lineals)
VII	ANTIQUA VARIANTS (ROMAN VARIANTS)
VIII	<i>Schreibschriften (Scripts)</i>
IX	Handschriftliche Antiqua (Graphics)
X	Fraktur (black letters)
	Xa Gotisch
	Xb Rundgotisch
	Xc Schwabacher
	Xd Fraktur
	Xe Fraktur Varianten
XI	Fremde Schriften (Non-Latin Scripts)

Figure 1: DIN 16518 type classification (with English translations taken from the British Standard 2961).

The St-G-typeface and DIN 16518

St-G-types (Figure 2) show influences from several scripts, forms of writing and writing systems: 1) from the *Akzidenz-Grotesk*, a typeface developed by the *Berthold AG* foundry that can be considered a ‘standard typeface’; 2) from book hand scripts such as the Roman Uncial and Carolingian Minuscule (i.e. <f>, <k>, <t>, <T>); and 3) from the Greek alphabet (i.e. <e>, <k>, <t>, <w>) (Lucius 2012).

Considering the linearity of the strokes and the absence of serifs, all St-G types seem to be contained by class VI, the sans serif typefaces. This assignment appears to fit properly for types such as <a>, <l> and <U>, formerly belonging to the *Akzidenz-Grotesk*, but it hardly seems suitable to put types with unusual shapes such as <A> into the same class. Do they rather belong to class VII, the Antiqua Variants? What about the types that seem to be based on historical book hand scripts, such as <e>, <t>, <w> and <k>? Do they have to be placed into class IX, the handwritten Antiqua? What if these types do not originate from book hand scripts but rather from the Greek alphabet? Should these types be put into class XI, the non-Latin scripts or maybe even into both classes? Ultimately, does it even make sense to assign a single type to four different classes?

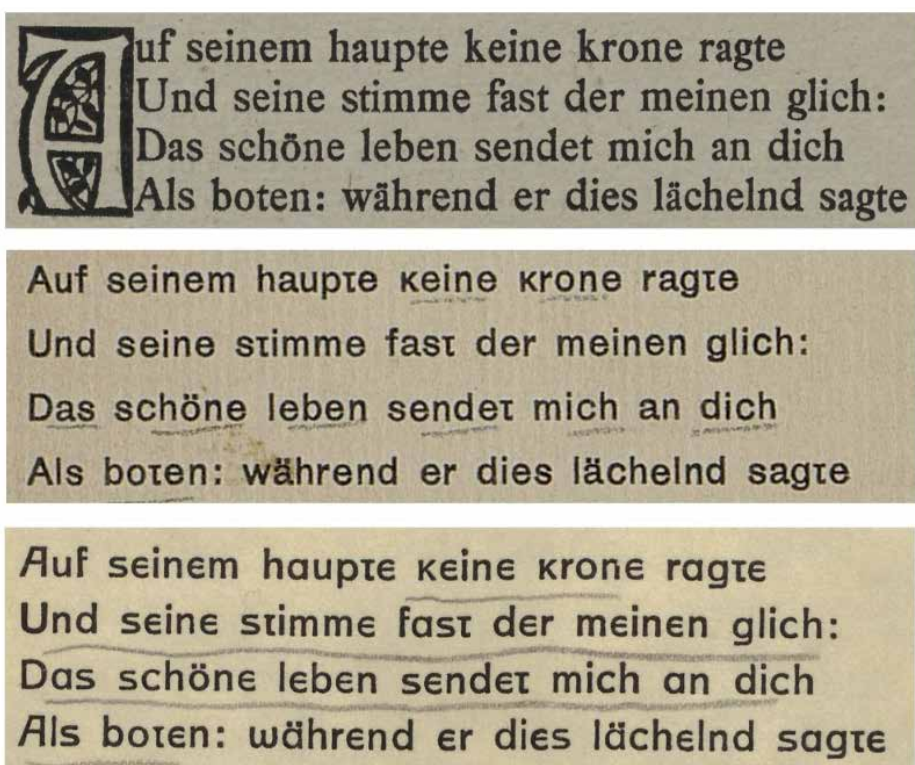


Figure 2: Sample poem of ‘Der Teppich des Lebens’ from different editions (from top to bottom): 1901, SBB-SPK Berlin Yo 28000; 1904, ULB Düsseldorf SPK DLIT8400; 1902, SBB-SPK Berlin Yc 11500-5.

Evaluating the DIN standard

The application of the DIN standard to the St-G-typeface raises more questions than it provides satisfying answers. The fact that St-G is actually not classifiable with the DIN norm underlines again the typeface’s exceptional nature, a feature that one recovers in many aspects of George’s poetical program. However, for the distinct identification of types and their placement into the history of typography within the context of a digital scholarly edition, the DIN standard is not very expressive because of several fuzzy aspects in its architecture, which may be summarized as follows:

- **Purpose and consistency:** the properties by which the type classes are defined are not consistent: some classes refer to a historical period (I-IV), others to forms (V-VII, X). Some classes are defined by the language or writing system (XI) and others again refer to a writing mode (VIII-IX). As a result, the purpose of the classification is not clearly comprehensible.
- **Terminology and semantics:** the class descriptions – in prose – are inexplicit: i.e. in the *Barock-Antiqua* class, the axis of the curves is described as ‘vertical or inclined slightly to the left’ (see British Standard). Vertical *or* slightly to the left and *how* slightly? Besides, the classification does not apply a shared vocabulary which increases terminological inaccuracy and especially impedes international communication.

- **Cross-classification and meta-classes:** classes I-V overlap regarding the properties ‘serifs’ and ‘stroke contrast’. Moreover, the seventh class ‘Antiqua Variants’ is entirely unspecific and does not provide any information about the objects it contains.

A key prerequisite for a sustainable classification is a well-defined domain and an explicit outline of the classification’s purpose in order to be able to comprehend the distinctive properties that were chosen. A valuable classification system should develop a stable terminology and a controlled vocabulary as well as apply formal modelling rather than modelling in prose. It should allow for an identification of concepts through unambiguous labeling of both multilingual variants and alternative terms for synonyms. Furthermore, it should avoid meta-classes, known as ‘other-classes’, since they run the risk of becoming too heterogeneous. Finally, if the classification lacks a hierarchy, defined by Sperberg-McQueen as ‘one-dimensional’, only one value per class should be assigned to avoid cross-classification (Drucker *et al.* 2014: ch. 2B; Sperberg-McQueen 2004: ch. 14).

Additional requirements

In addition to the improvements suggested above, a sustainable type classification should reflect that type design is still a growing field and that exceptions such as the St-G-typeface always will exist. Hence, it is impossible to pin down the content of classes into hard and fast definitions and properties need to be liberated from their dependency on classes. Rather than predefine areas of knowledge and carrying out a top-down classification, classes should be pieced together through the combination of properties and as such bottom-up. Furthermore, to assure a certain degree of granularity, the classification should imply both hierarchical and associative relationships (Balzer and Lindenthal 2013: 3-5).

Another important point to consider when building a type classification is the wide range of properties according to which typography can be classified: a graphic designer might want to put the focus on properties of aesthetics and manufacturing while a literary scholar might be interested in semiotic aspects. A type classification in the context of the Stefan George edition requires the consideration of ‘form’ and ‘style’ as well as the inclusion of parameters, which express ‘function’ (i.e. headline, paragraph, emphasized text parts), and ‘emotional impact’ (i.e. static, heavy). In order to reflect various aspects of typography, the properties could be organized for instance in multiple hierarchies. Each tree could reflect a different typographical dimension and they could be combined flexibly which each other.

Conclusion

The failed classification of St-G with the DIN 16518 standard again emphasizes the unconventional design of the typeface. However, it is of little help to identify and classify the types in the context of the digital George edition. Therefore, it cannot serve as a basis for an examination of the poetical function of typography in Stefan George’s work. Even if it is not likely that a single classification scheme is exhaustive or can meet all needs, it is evident that the DIN standard is in need of

revision since it is not consistent, terminologically sharp nor extensible enough to identify types and classify them in an unambiguous way.

Considering the established requirements for the sufficient reflection on the particularities of St-G, it is discussable whether a standard classification structure is powerful enough to meet all the outlined features or if ontologies might be a better solution. Classifications and ontologies overlap greatly and follow very similar principles. However, ontologies, with their stronger potential to represent network-like structures, appear to be more suitable to guarantee flexibility concerning the combination of properties, extensibility through the introduction of new properties, and multidimensionality of the typographical contents. Only by including these key functions, a valuable classification of St-G-types as well as a subsequent analysis of their poetical function can be assured for the Stefan George edition.

References

- Balzer, Detlev and Jutta Lindenthal. 2013. 'Building better controlled vocabularies: guidance from ISO 25964', *presentation at ISKO UK 2013*, <http://www.isko-uk.org/content/building-better-controlled-vocabularies-guidance-iso-25964>.
- Drucker, Johanna, Kim David, Iman Salehian and Anthony Bushong. 2014. 'Classification Systems and Theories' (ch. 2B), *Introduction to Digital Humanities*, <http://dh101.humanities.ucla.edu/>.
- Kupferschmid, Indra. 2012. 'Type classifications are useful, but the common ones are not.' *Kupferschrift* (blog), March 31, 2012, <http://kupferschrift.de/cms/2012/03/on-classifications/>.
- Runk, Claudia. 2006. *Grundkurs Typografie und Layout*. Bonn: Galileo Press.
- Sperberg-McQueen, C. Michael. 2004. 'Classification and its Structures' (Ch. 14), *Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, <http://www.digitalhumanities.org/companion/>.
- Wolf, Lucius D von. 2012. 'Buchgestaltung und Typographie bei Stefan George' (Ch. 5. 6; p. 467-491), *Stefan George und sein Kreis, Ein Handbuch*, vol. 1, edited by Achim Aurnhammer, Wolfgang Braungart, Stefan Breuer, Ute Oelmann. Berlin De Gruyter.

Visualizing collation results

Elisa Nury¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

In the recent years, the use of automatic collation tools such as CollateX has increased, and so has the need to display results in a meaningful way. Collation is admittedly more than just a record of variant readings (Macé *et al.* 2015, 331). It frequently incorporates additional notes and comments, 'paratextual' elements such as changes of pages or folia, gaps, lacunae, and so on. This combination of variants and paratextual material produces a large amount of complex collation data, which are difficult to read and interpret. Therefore, the editor needs to visualize and analyse the collation results as a whole, and not only variant by variant. A good visualization should offer a way to check collation against the actual witnesses, whether they are manuscripts or printed editions. In addition, the editor should be able to interact with the collation to analyse readings and variants. Collation could be filtered, so as to find patterns of agreements or disagreements between those witnesses, which can indicate how they are related to each other. Visualization and manipulation of collation results are thus essential in order to use collation for further research, such as studying the manuscript tradition and creating a *stemma codicum*.

To tackle these issues faced by an editor, I would like to present a method of visualization of collation results. The method of visualization consists of two aspects: first, a description of the collation table displayed in HTML; second, a Jupyter notebook² where the editor can interact with the collation through a Python script, for instance to select agreements between a group of witnesses against another group, or to make small corrections in the alignment. The case study to which the visualization was applied is the *Declamations* of Calpurnius Flaccus. It is a classical literary text in Latin from the second century. The witnesses, four manuscripts and two critical editions, have been encoded in TEI P5 and then pre-tokenized for collation with CollateX into a JSON format that allows to record a reading together with more detailed information.

1 King's College London; elisa.nury@kcl.ac.uk.

2 <https://jupyter.org/> (Accessed November 8, 2016).

The Collation Table

The collation table is a visualization that is user-friendly for scholars who do not work with CollateX or any computer-supported collation program. The table typically represents each witness on a separate line or column, with their text aligned when it matches, and blank spaces inserted where a part of the text is missing in a witness. In its most simple form, the table will show only plain text from the witnesses. Enhancements have been proposed to improve this basic table, for instance with colours to indicate the places where a variation occurs: in the Digital Mishnah demo, variant locations are highlighted in grey.³ Another example is the Beckett Archive project, where deletions are represented with strikethrough and additions with superscript letters.⁴ However, other elements are still missing from those helpful visualizations, such as, for instance, the changes of folia mentioned earlier. The reason for recording folia changes is mainly for checking purposes. If the editor or a reader wants to check the accuracy of the transcription for a particular reading, it will be much easier to find the reading back in the manuscript knowing the folio where it appears. How could this or similar paratextual elements be integrated into collation results? One solution is to take advantage of the JSON input format of CollateX.

By default, CollateX can take as input plain text transcriptions of the witnesses to collate. The texts will be split into ‘tokens’, smaller units of text, at whitespaces. This is the tokenization stage.⁵ The collation is then performed on these tokens, which are usually the words of the text. However, it is also possible to ‘pre-tokenize’ the transcriptions. The JSON format, in particular, allows to record not only the plain text words (t), but also other properties, such as a normalized form of the word (n). There is no limitation to the token properties that can be added: they simply will be ignored during the collation stage, but still be available in the end results. In order to integrate folio location, links to digital images and editorial comments, I transformed the XML TEI transcriptions of Calpurnius Flaccus into pre-tokenized JSON. The tokens include, beside the (t) and (n) properties, a (location) property, eventually a (link) and/or a (note) property. Below is an example of a collation table for the *Declamations* of Calpurnius Flaccus, making use of those properties. I have created this table by comparing the agreements of normalized readings among the different witnesses, with the help of the Jupyter notebook described below.

There are four manuscripts in this collation: B, C, M and N. Each manuscript is divided into two witnesses according to the different hands that wrote the text. For example, B1 is the first hand of manuscript B and B2 is the second hand who made corrections to the text of the first hand. There are also two editions in the collation. The *editio princeps* was first published in 1580 by the French scholar Pierre Pithou and reprinted in 1594. The second is the critical edition of Calpurnius Flaccus published by Lennart Håkanson in 1978. The last column, ID, represents a way to identify rows in the table. There are a few items highlighted in the table of Figure 1:

3 See the demo here: <http://www.digitalmishnah.umd.edu/demo> (Accessed November 4, 2016).

4 See the news update of September 17, 2014: <http://www.beckettarchive.org/news.jsp> (Accessed November 4, 2016).

5 See CollateX documentation: <http://collatex.net/doc/#input> (Accessed November 4, 2016).

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
excerpta	excerpta	excerpta	excerpta							1
					contra matrem	contra matrem	contra matrem	contra matrem	contra matrem	16
					contra matronam	contra matronam	contra matronam	contra matronam	contra matronam	24
					pro milite	pro milite	pro milite	pro milite	pro milite	65
00	00	00	00		0	0	0	0		67
Note: Unknown abbreviation. Normalized form supplied by Lehnert.	148r:8	82r:18	82r:18		2r:7	2r:7	244v:28	244v:28		
148r:8										
verginus	verginus	verginus	verginus	Verginius	virginus	virginus	virginus	virginus	Virginius	76
					pro parricida	pro parricida	pro parricida	pro parricida	pro parricida	84
pater	pater	pater	pater	patiar includi	paterer	paterer	paterer	paterer	patiar includi	124
est	est	est	est	es	es	es	es	es	es	127

Figure 1: collation table extract.

- The (i) symbol next to a reading: on click, it can reveal/hide editorial comments that were made during the transcription, especially regarding problematic passages. Here the comment is related to an unknown abbreviation that was not resolved with certainty by the editors of Calpurnius.
- The (:) symbol in the ID column: on click, it will reveal/hide locations of the readings for each witness in the row.
- The location is in the form of ‘folio number:line number’. The unknown abbreviation mentioned above appears in folio 148r, line 8, of manuscript B. Since there is a digital facsimile available for manuscript B, the location will also link to the image of the page.
- Green and red colours: the coloured lines next to a reading show agreement (green) or disagreement (red) with the same reading of a base text, chosen among the witnesses. Here the base text is the text printed in the edition of Håkanson. As a result, the readings Håkanson rejected because he considered them to be errors are shown in red, while the reading he accepted as true are shown in green. The pattern of colours would of course be different if another text, such as Pithou’s edition, had been selected as the base text.

The purpose of the colours is to detect relationships between witnesses, according to the (neo) Lachmannian⁶ method of text editing. Lachmann’s method focuses on common errors shared by a group witnesses in order to postulate relationships between those witnesses: a group of witnesses are likely to be related if they (1) agree on readings that (2) they do not share with the other witnesses, and especially (3) when they agree in errors, i.e. when they share readings that have no manuscript authority⁷ and do not represent the original text. Using the red/green colour scheme was inspired in part by Stemmaweb, a tool for creating

⁶ Here neo-Lachmannism refers to the improvements to Lachmann’s method brought by Pasquali and other Italian scholars, who took Bédier’s criticism into account and incorporated the study of the textual tradition and material documents (the manuscripts themselves) to the creation of stemmata. In this sense, neo-Lachmannism is also based on common errors shared by witnesses.

⁷ A reading with manuscript authority is ‘a reading that may have reached us through a continuous sequence of accurate copies of what the author wrote back in antiquity and may therefore be authentic and (by definition) right’ (Damon 2016, 202-203).

stemmata with computer algorithms (Andrews 2012).⁸ Being able to find common errors in the collation results would be especially useful for a scholar preparing a critical edition. For this purpose, I have prepared a python script, in a Jupyter notebook, which lets users filter the collation in order to find witnesses that agree with each other, and not with others.

The Jupyter Notebook

The Jupyter notebook offers an interactive way to explore collation results thanks to widgets, which are components of a user interface such as buttons, textboxes, and so on. Here I will show only one example of interaction, the selection of agreements between witnesses (Figure 2).

This small extract from the Jupyter notebook shows two selection widgets in the form of dropdown menus. In the first menu, the user selects a list of witnesses to see where they agree with each other. In the second menu, the user can select another list of witnesses. The resulting collation table will show readings where the first witnesses agree with each other and not with the witnesses in the second list.

The selection of witnesses in this example will result in the collation table that was presented in Figure 1. The table shows the agreements of B and C (both hands) against readings of M, N and P1594. This does not mean that M, N and P1594

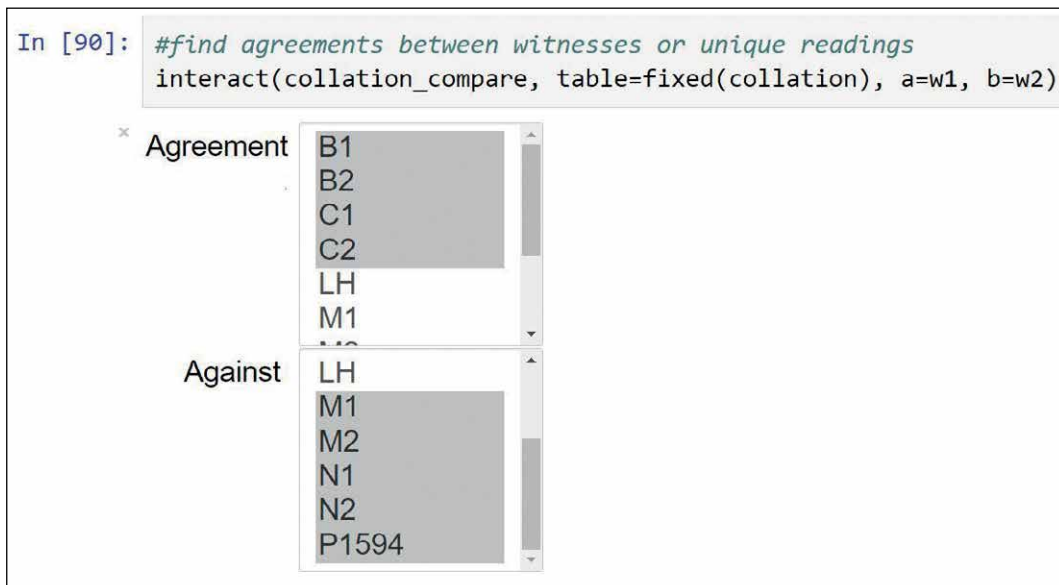


Figure 2: widgets from the Jupyter notebook.

8 Stemweb makes it possible to visualize collation tables where readings are highlighted in green or red. A green highlight means that the reading is consistent with a given stemma hypothesis. A red highlight means that the reading is not consistent with that same stemma hypothesis. In a similar way, the collation table highlights readings that are consistent or not with a 'text hypothesis,' the text that was selected as a base text.

agree together, only that they are different from B and C for the readings displayed in the table. The full table shows many examples of B and C agreeing together, and especially agreeing in errors. It shows that, at least according to Håkanson, they are related witnesses. In fact, editors of Calpurnius Flaccus recognize that B and C form a closer group, while M and N form another group of manuscripts (Sussman 1994, 19). The *editio princeps* of Pithou, on the other hand, is believed to be related to manuscript N.

The notebook offers more interactions with the collation results. Beside searching for agreements, it is possible to modify the witnesses' alignment, add notes to readings or search for a specific reading. Although the rest of the notebook is not discussed here, the code is made entirely available on Github.⁹

In conclusion, the two examples of visualization described in this paper demonstrate how to make use of collation in an electronic format for further research. CollateX results can be improved with the use of pre-tokenized JSON, as it was already done by other projects such as the Beckett Archive. However, it is possible to integrate more information into the collation with JSON tokens: elements such as location of a word in the manuscript, or editorial comments, are important aspects of collating texts and there is no reason to discard them in a computer-supported collation. As shown in the collation table, the use of a few symbols allows one to make those elements easily available without overcrowding the results. The use of colours is a straightforward way to reveal groups of witnesses that agree with one another and thus help draw conclusions about the manuscript tradition. The collation table and Jupyter notebook presented here hopefully will provide suggestions on how to make available the extra material that is not yet exploited fully in collation visualizations.

References

- Andrews, Tara. 2012. 'Stemmaweb – A Collection of Tools for Analysis of Collated Texts.' <https://stemmaweb.net>.
- Damon, Cynthia. 2016. 'Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts.' In *Digital Scholarly Editing. Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 201-218. Open Book Publisher.
- Macé, Caroline, Alessandro Bausi, Johannes den Heijer, Jost Gippert, Paolo La Spisa, Alessandro Mengozzi, Sébastien Moureau and Sels Lara. 2015. 'Textual criticism and text editing.' In *Comparative Oriental Manuscript Studies: An Introduction*, 321-466. Hamburg: Tredition. <http://www1.uni-hamburg.de/www/COMST/comsthandbook/321-466> Chapter 3.pdf.
- Sussman, Lewis A. 1994. *The Declamations of Calpurnius Flaccus: Text, Translation, and Commentary*, edited by Lewis A. Sussman. Mnemosyne Bibliotheca Classica Batava. Leiden: New York: E. J. Brill.

9 <https://github.com/enury/collation-viz> (Accessed November 3, 2016).

The Hebrew Bible as data: text and annotations

Dirk Roorda¹ & Wido van Peursen²

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

The Eep Talstra Centre for Bible and Computer (ETCBC) (1) is specialized in the study of the Hebrew Bible. Its main research themes are linguistic variation in the historical corpus, identification of the linguistic system of Hebrew and the way it is used in narrative, discursive and poetic genres, and solving interpretation riddles using thorough data analysis.

To this end, the ETCBC has created a text database of the Hebrew Bible, annotated with linguistic information. This database harbours decades of encoding work.

In 2014, the CLARIN-NL (7) project SHEBANQ (2) has made this work available online as the website SHEBANQ. A powerful tool, LAF-Fabric, also was developed for manipulating this data. Roorda (2015) gives a concise history of this database and a description of SHEBANQ.

In the present article we focus on the methodological choices that eventually helped Hebrew text processing along. Three elements are key: an abstract text model, separation of concerns, and performance. We also comment on the sustainability of this work.

1 dirk.roorda@dans.knaw.nl.

2 w.t.van.peursen@vu.nl.

Abstract text model

Consider the first verse in de Hebrew Bible, Genesis 1:1 ('In the beginning God created the heavens and the earth'). If you look it up in a standard text, such as the Biblia Hebraica Stuttgartensia (Elliger and Rudolph 1997), you see Figure 1.

בְּרֵאשִׁית בָּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ:

Figure 1.

The diacritics represent vowels and accents, and the big, square letters are the consonants. The Hebrew language is such that to native speakers the text is perfectly readable without the vowels, and that is the usual way of spelling up till now. In the Hebrew Bible the consonants are many centuries older than the vowels!

בראשית ברא אלהים את השמים ואת הארץ:

Figure 2.

Hebrew is written from right to left, which poses problems for various applications, especially where Latin text is mixed with Hebrew text, as in syntax trees and basically everywhere where linguistic observations are being made. That means that it is convenient to have a faithful ASCII transliteration of the material, such as the in-house spelling at the ETCBC (Figure 3).

B.:- R;>CI73JT B.@R@74> >:ELOHI92JM >;71T HA- C.@MA73JIM W:- >;71T H@- >@75REY00

Figure 3.

For yet other applications it is handy to have a phonetic representation of the text, *e.g.* Figure 4.

b^ərēš^hīt bār'ā ʔ^əlōh'īm ʔ^əēt haššām'ayim w^ʔēēt hāʔāreš .

Figure 4.

We have at least five ways to code the text and yet we do not want to consider the resulting text strings as five different texts, but as five representations of the same text. That means that our model of text cannot coincide with the very concrete notion of a string of characters. We need more abstraction.

Another reason for more abstraction is that not all words are separated by whitespace. Some words attach to adjacent words. We need a word distinction that is more abstract than the obvious, concrete word distinction by whitespace.

Finally, the text is segmented. There are books, chapters, verses, half-verses, sentences, clauses, phrases, subphrases and words. Segments can be embedded in other segments. Some segments become interrupted by other segments. The segmentation and embedding are objective characteristics of the text, and we want to represent those in our text model, despite the fact that much of it is not marked explicitly in the text string.

We are not the first to explore abstract text models. In his PhD thesis Doedens (1994) provides exactly the model that in subsequent years spawned a text database system by Petersen (2004) and a practice of translating exegetical problems into syntactical queries on the text as data(base) (Talstra and Sikkel 2000).

The essence of this model is this: one chooses a basic unit of granularity, such as character, morpheme, or word (in our case: word), and considers them as numbered objects. The numbering corresponds to the sequential order of these objects, which also are called *monads*. All other objects are subsets of the set of all monads. One object is embedded in another one, if the monads of the first object form a subset of the monads of the second object. All objects can have features. Features are basically key-value pairs, such as *part-of-speech* = *noun*, or *consonantal-translit* = ‘B>R>’. The model organizes the objects into types, where the type determines which features its objects can have. We have a type for phrases, which is different from a type for clauses, because our encoding has bestowed different kinds of features on phrases than on clauses.

Summarizing: our abstract text model represents the text as a database of objects, with a notion of sequence and embedding. Objects can have any number of features, where features are organized in object types.

We want to stress that such an abstract text model moves away from regarding a text as coincidental with any representation it has. The text is not the one- or two-dimensional collection of glyphs on a canvas. It is rather a family of objects with spatial relationships and representational and functional properties. The actually given, concrete text can be obtained from it, but the abstract model is richer, because many other representations can be derived from it as well.

An other point to note is that the notion of embedding in the model by no means implies that the text is organized in a single, comprehensive hierarchy. The model is much more liberal: some parts may be outside any hierarchies, there might be multiple hierarchies and those hierarchies might overlap in arbitrary ways. This is a good thing, because in our case our data exhibits most of these phenomena.

This all might look pretty well from the theoretical perspective, but is it practical? Are there standards for this type of encoding? And what about TEI?

The good news is, that there is a standard that comes very close to capturing the ideas sketched above. It is Linguistic Annotation Framework (LAF) (3) (Ide and Romary 2012). In LAF, there is a graph of nodes and edges, where the nodes correspond to regions of primary data. Nodes and edges can be annotated by feature structures. The correspondence with the previous model is apparent: nodes can be used to represent objects, and feature structures are a generalization of plain features. Further, in LAF, there is not a built-in concept of sequence and embedding. Instead, there are the regions as a category between the primary data

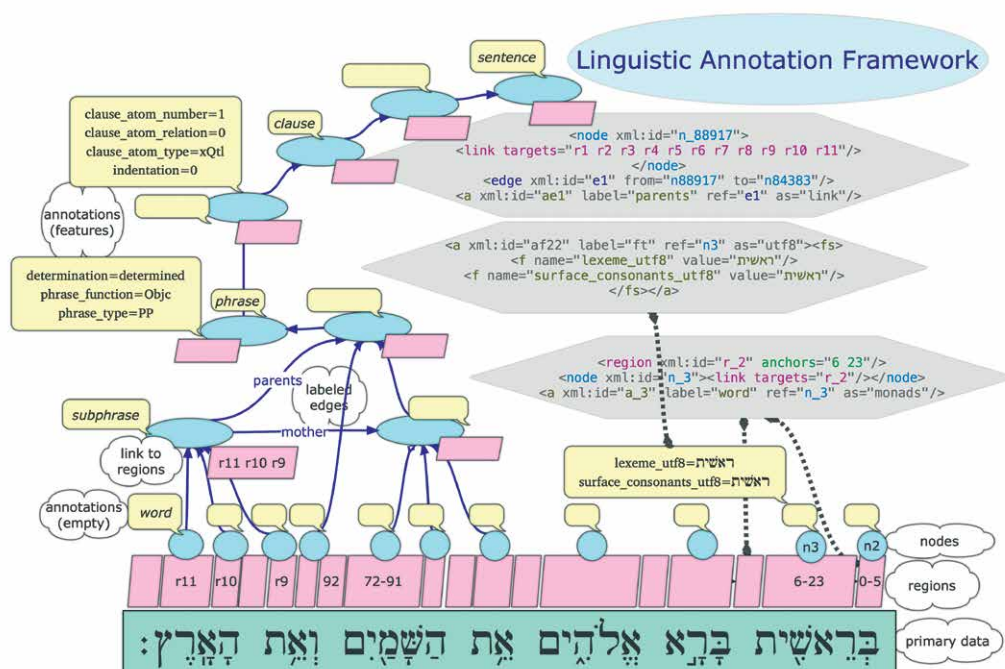


Figure 5.

and the nodes. By defining your regions, you can get the notion of sequence back. The edges can be used to relate nodes to each other. The model allows multiple, annotated relations. In practice, unlabelled edges are used to model the embedding relationship.

Another piece of good news is that the whole LAF model can be coded into XML. The encoding scheme is very light-weight: most of the rules are about how to describe/declare the metadata. The model itself is coded in very simple tags for nodes, edges and annotations, where the linking is achieved by using XML attributes. The feature structures are coded according to the TEI module for feature structures (12).

Separation of concerns

When the Hebrew Bible is encoded in LAF, it is striking that the plain text is only 5 MB, whereas the annotations with the linguistic features take up 1.5 GB. Moreover, there are research projects under way that will deliver more annotations, some syntactical, some at the discourse level. There is a need for more semantic information, *e.g.* named entities and multiple senses of words. A separate, but not disconnected enterprise is the discipline of textual criticism, which leads to a mass of yet another type of annotation (see *e.g.* Rezetko and Naaijer 2016).

Last but not least, we also want to elicit user-contributed annotations, either in the form of manually added annotations, or in the form of bulk-uploaded software-generated sets of annotations.

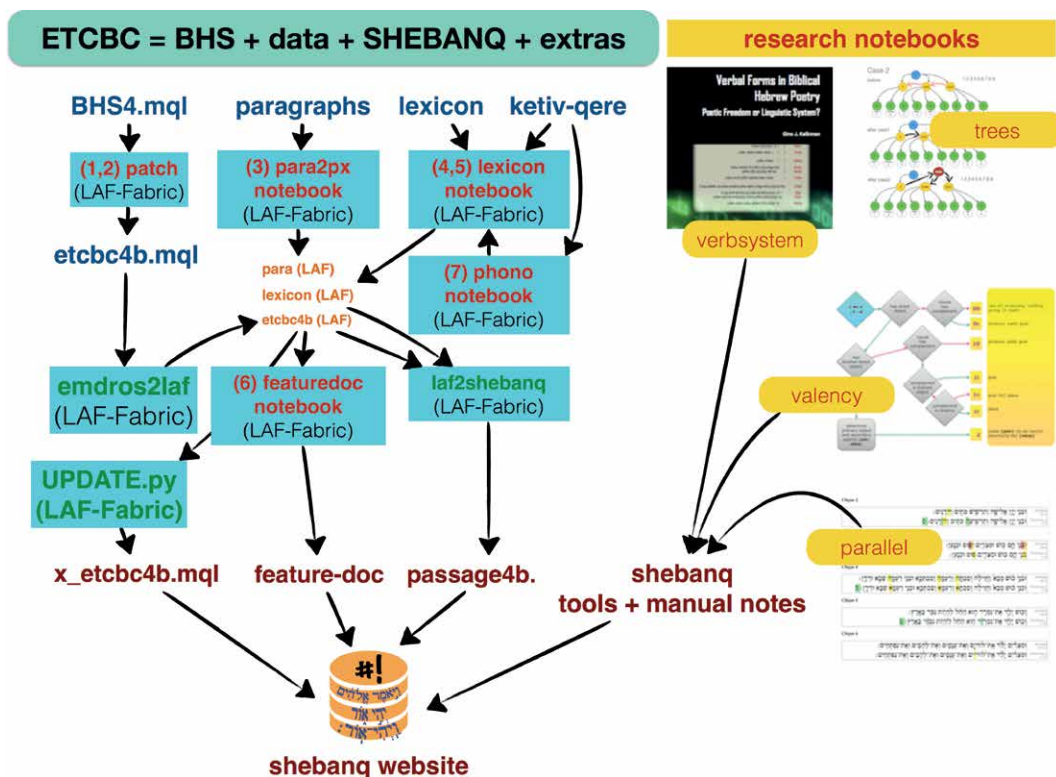


Figure 6.

These ambitions quickly will lead to unmanageable complexity of data and tools if our text model does not support the separation of concerns. The concept of primary data plus annotations already makes a very useful, principled distinction between an essential core of given and managed data on the one hand, and interpreted, reader-contributed data on the other hand. In order to create an annotation, there is no need to modify the primary data, not even for the purpose of inserting anchors.

The second separation is between the different sets of annotations. It is possible to conduct several annotation producing workflows at the same time, without the constant need to agree on a large set of encoding details.

The reader may have noticed between the lines: LAF is a strictly stand-off encoding practice. When we designed SHEBANQ in the course of a CLARIN-NL project we had to choose a standard format for the data. LAF seemed a good fit, and in the past years we have not regretted it and benefited from its stand-off character. The SHEBANQ tools page (5) gives a good indication of how new annotations and representations are going to live together in the whole SHEBANQ.

Performance

In order to process text, it has to be represented. In computing, part of the problem solving is finding a representation of the material that helps facilitate the tasks to be performed. Different tasks lead to different representations, and a lot of computing consists of moving information from one representation to another. SHEBANQ is a good example of this. Despite the suitability of LAF for representing the text, we need to transform the data again to process the material efficiently.

The Hebrew Bible in LAF is a massive resource: 1.5 GB of XML coding a web of objects with 30 million features. There are many tools for XML processing, but neither seemed to do a good job on this resource. An underlying reason is that our LAF is not document-like (a tree of deeply nested elements) and not database-like (a long list of flat records). It is graph-like, and the normal XML tools are just not optimized for it.

Our purpose has been to create a workbench where programming theologians can write simple scripts to walk over the data, grab the details they want, and save them in csv files in order to do interesting things with them, such as statistical analysis in R. We wrote a new tool, LAF-Fabric (4) that does exactly that. When first run, it compiles the LAF XML source into convenient Python data structures, and saves them to disk in a compact format. This may take 15 minutes. In all subsequent runs, the data will be loaded in a matter of seconds, so that the computer will work hard for the research task at hand, and not for the XML bureaucracy of parsing. Moreover, if you run your task in a Jupyter (13) notebook, you can load the data once, and develop your task without the need to repeat the data loading step all the time. This makes for very pleasant and efficient explorative data analysis.

As an example, the snippet of Python code in Figure 7 counts how many times each type of phrase occurs with outcomes (only the top 4) in Figure 8.

```
In [10]: phrase_types = collections.Counter()
         for node in F.otype.s('phrase'):
             phrase_types[F.typ.v(node)] +=1
```

Figure 7.

VP	68941
PP	57478
CP	52458
NP	40856

Figure 8.

LAF-Fabric is not only useful for counting stuff, you also can use it to produce texts in fine typographical output, visualizations, and forms to add new annotations and process them. The trade off is that you have to program it yourself. LAF-Fabric eases the programming only in the aspects that deal with the particularities of LAF. That makes it a suitable tool for data preprocessing as part of datamining or visualization, where the users are programmers themselves.

Recently, we exported the data to R format (.rds), resulting in a 45 MB table with all nodes and features in it (14).

Users now can employ the whole power house of R to explore the Hebrew Bible as data. In the same Jupyter notebooks in which they perform data preprocessing with LAF-Fabric, they also can produce R analysis and graphics. There are several introductions to R for humanists and linguists: Baayen 2008, Levshina 2015, and Arnold and Tilton 2015.

Sustainability

SHEBANQ is a system that relies heavily on a particular dataset and on particular software tools. So how do we sustain this web of machinery? How can a research institute bear a burden like this?

Part of the answer is that a single, isolated research group will have a hard time to keep SHEBANQ alive. In our case, we are helped by supporting institutions (academic repositories) and global services (Github, Zenodo).

Data

Our data is archived at DANS (6), which takes care that the ETCBC data is long-term archived, findable, referable, downloadable and usable (Van Peursen *et al.* 2015).

The most recent version of the data is also on Github (8), as a convenient service to people who want to get started quickly with LAF-Fabric. Previous versions also are kept at DANS, and inside SHEBANQ, since there are published queries that depend on it.

Note that it is possible to reconstruct the plain text of the Biblia Hebraica Stuttgartensia from the data. There is a Creative Commons Attribution Non-commercial license on this data, in agreement with the publisher of the BHS, the German Bible Society (9).

Software

Software is much more difficult to sustain. The difference with data is that data should be kept constant over the years, whereas software has to evolve with the flow of technology in order to remain usable.

There is no perfect solution to software sustainability. What we have done is a compromise.

The query engine behind SHEBANQ is Emdros (Petersen 2004). We do not preserve the engine indefinitely, but we preserve the results obtained with it. We record when a query has been written, executed, and by which version of Emdros. If the time comes that we no longer can employ the Emdros software, we still can

point at the results of the queries in SHEBANQ, some of which may have been cited in publications.

The underlying software for SHEBANQ and LAF-Fabric is on Github, open source. There are snapshot versions archived at Zenodo. The repositories contain their own documentation. From the start of 2017 onwards, I have deprecated LAF-Fabric in favour of a new format and tool: Text-Fabric.

General

We think we are well-positioned to preserve the results obtained by SHEBANQ for a long time. We will not be able to preserve the software indefinitely, but at least that will not break the preservation of the data, nor the interpretation of the results that we do preserve.

The best guarantee for sustainability might be not mere preservation, but active dissemination. Only when enough people use the system and benefit from it, is there a chance that it survives the first decade.

We are pleased to observe that several other groups are looking at the ETCBC data and using it to create their own systems. The next step is to link those systems in meaningful ways, so that researchers can hop around easily for the specific functionality they need. A good example is the Bible Online Learner (Winther-Nielsen and Tøndering 2013) with links to SHEBANQ and vice versa.

References

- Arnold, Taylor and Lauren Tilton. 2015. *Humanities Data in R. Quantitative Methods in the Humanities and Social Sciences*. Springer. With supplementary data. <http://humanitiesdata.org/>.
- Baayen, Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics*. Cambridge University Press. Draft available as PDF: <http://www.sfs.uni-tuebingen.de/de/~hbaayen/publications/baayenCUPstats.pdf>.
- Doedens, Crist-Jan. 1994. *Text Databases. One Database Model and Several Retrieval Languages*. Number 14 in Language and Computers, Editions Rodopi, Amsterdam, Netherlands and Atlanta, USA. ISBN: 90-5183-729-1, <http://books.google.nl/books?id=9ggOBRz1dO4C>.
- Elliger, Karl and Wilhelm Rudolph (eds). 1997 (5th corrected edition). *Biblia Hebraica Stuttgartensia*. Stuttgart: Deutsche Bibelgesellschaft. <http://www.bibelwissenschaft.de/startseite/wissenschaftliche-bibelausgaben/biblia-hebraica/bhs/>
- Ide, Nancy and Laurent Romary. 2012. *Linguistic Annotation Framework. ISO standard 24612:2012*. Edition 1, 2012-06-15. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=37326.
- Levshina, Natalia. 2015. *How to do Linguistics with R. John Benjamins publishing*. <https://benjamins.com/catalog/books/z.195/main>.
- Petersen, Ulrik. 2004. 'Emdros – A Text Database Engine for Analyzed or Annotated Text'. In *Proceedings of COLING*: p. 1190-1193. <http://emdros.org/>.
- Rezetko, Robert and Martijn Naeijer. 2016. 'An Alternative Approach to the Lexicon of Late Biblical Hebrew.' *Journal of Hebrew Scriptures* (jhsonline.org).

- Roorda, Dirk. 2015. The Hebrew Bible as Data: Laboratory Sharing – Experiences. arXiv:1501.01866.
- Talstra, Eep and Constantijn J. Sikkel. 2000. 'Genese und Kategorienentwicklung der WIVU-Datenbank.' In *Fontes! Quellen erfassen lesen deuten. Wat ist Computerphilologie? Ansatzpunkte und Methodologie Instrument und Praxis*, edited by C. Hardmeier *et al.* Amsterdam: Ad VU University Press: pp. 33-68.
- Van Peursen, Wido, Constantijn Sikkel and Dirk Roorda. 2015. Hebrew Tekst Database ETCBC4b. Dataset archived at DANS. DOI: 10.17026/dans-z6y-skyh
- Winther-Nielsen, Nicolai and Claus Tøndering. 2013. Bible Online Learner. Website, learning system for Hebrew students. <http://bibleol.3bmoodle.dk/>.

- (1) <http://www.godgeleerdheid.vu.nl/en/research/institutes-and-centres/eep-talstra-centre-for-bible-and-computer/index.asp>
- (2) <https://shebanq.ancient-data.org/>
- (3) http://www.iso.org/iso/catalogue_detail.htm?csnumber=37326
- (4) <https://github.com/Dans-labs/text-fabric/wiki>
- (5) <https://shebanq.ancient-data.org/tools>
- (6) <http://www.dans.knaw.nl/en>
- (7) <http://www.clarin.nl/>
- (8) <https://github.com/ETCBC/bhsa>
- (9) <https://shebanq.ancient-data.org/sources>
- (10) SHEBANQ snapshot: <http://dx.doi.org/10.5281/zenodo.33091>
- (11) LAF-Fabric snapshot: <http://dx.doi.org/10.5281/zenodo.33093>
- (12) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html>
- (13) <http://jupyter.org/>
- (14) <https://shebanq.ancient-data.org/tools?goto=r>

Full Dublin-Core Jacket

The constraints and rewards of managing a growing collection of sources on omeka.net

Felicia Roşu¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

The best digital editions of correspondence available today – the Van Gogh Letters Project, Mapping the Republic of Letters, Electronic Enlightenment – are true flagships for the current state of the art in digital humanities. But I would like to draw attention to a different type of project, more modest but at the same time more accessible to non-DH specialists who may contemplate digital source publication: an Omeka-based, almost DIY document collection that is expanding constantly and has evolving metadata that fits imperfectly within the Dublin Core element set. My paper reflects on the advantages and disadvantages of omeka.net and it will offer a historian's 'lay' perspective on the constraints and rewards of using controlled vocabularies, CSV imports, and Dublin Core elements for research and teaching purposes.

Our document collection (Vincentian Missionaries in Seventeenth-Century Europe and Africa: A Digital Edition of Sources²) began in an XML and TEI environment, which was handled for us by the enthusiastic team of the now-closed Digital Humanities Observatory (Dublin, Ireland). When our initial funding period ended, the digital edition – which was only an optional component of a larger research project led by Dr. Alison Forrestal at the National University of Ireland, Galway – was far from completed. There were some nice spatial and

1 f.rosu@hum.leidenuniv.nl.

2 <<http://earlymoderndocs.omeka.net/>>.

temporal visualizations, but our items did not have any attached transcriptions and the metadata itself needed a lot more work.

Two years later, we decided to revive the database with the help of a new sponsoring institution (DePaul University from Chicago), which had a keen interest in our topic but could only offer limited funding. Their Digital Services staff recommended moving our data to omeka.net because of the low maintenance, user friendliness, and export capabilities of that platform. The new funder agreed to pay for an annual subscription to the Omeka Gold package; their Digital Services staff introduced us to the platform and uploaded a trial version of the collection; and later they also provided occasional but crucially important advice for the problems that arose from our need for periodic updates. But mostly, after the learning phase, we were left on our own.

Omeka.net offers limited customization and visualization options, but it is a growing platform with an increasing number of plugins; most importantly, it does not require heavy technical assistance. I liked the arrangement because handling the collection was not my main research activity and Omeka offered the flexibility and freedom of a true passion project – one on which I worked in my free time, with only one student assistant. Even with such limited commitment, the collection grew exponentially. Two years ago, the online edition had 120 items and no attached transcriptions. It now has 690 items, 540 attached transcriptions, increasingly complex metadata, and it continues to expand. Last but not least, it inspired my first student assistant (Thérèse Peeters, Leiden University) to start a PhD on a topic derived from its documents, and it continues to be a valuable teaching resource.

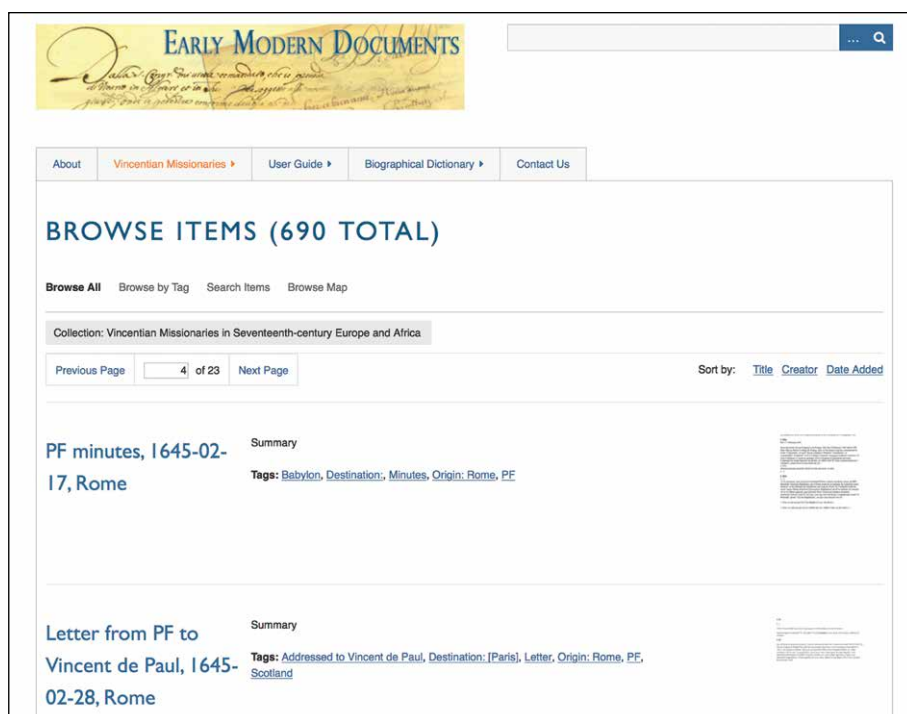


Figure 1.

The challenge of repeated CSV imports: how we dealt with diacritics, attachments & geolocation

Instead of waiting for years until our work was completed, we decided to make it public as soon as possible and then update it as we went along. However, online edits on omeka.net do not work well for us because they can only be made on an item-by-item basis and we often make spelling or structural changes across the board, as our work with the sources evolves. For that reason, we keep our metadata in a spreadsheet that is constantly modified offline. Periodically, we convert it to CSV and upload it onto Omeka via the CSV import option offered by the platform.

Diacritics were one of the first problems we encountered in the uploading process. The problem came from Excel, which does not create UTF-8 formatted text fields in their spreadsheets. The digital specialists from DePaul University advised us to import our spreadsheet into Google Docs, because Google spreadsheets are UTF-8 formatted. Indeed, when we exported CSV files from Google Docs, we were able to retain the diacritics.

Attaching transcriptions was another problem. Initially we thought that we could only attach transcriptions online, item by item, via Omeka’s ‘File’ option. We did not mind doing that once, but uploading new versions of the database meant that the attachment operation had to be repeated all over again. Luckily, our colleagues from DePaul found a solution for us. We placed our transcriptions in a public Dropbox folder (for some reason, Google Drive did not work); then, in our spreadsheet, we added a column containing their permanent URLs; and finally, in the CSV import operation, we identified that column as ‘files’ (see below). We may now modify the metadata as often as we wish, without having to reattach the transcriptions with each new upload of the spreadsheet (Figure 2).

Unfortunately, we found no solution for geolocation, which can only be added separately. This is why for the moment we only have visualizations for 115 of our items, which were mapped by our collaborator, Niall O’Leary, in an earlier stage of the project.³ I hesitate to tell Niall to map all of the 690 items we currently have online when I know we are preparing the next upload, so we will leave geolocation for the very last stage of our project – unless Omeka introduces geolocation among its CSV import options.

Step 2: Map Columns To Elements, Tags, or Files					
	Example from CSV File	Map To Element	Use HTML?	Tags?	Files?
Attachment	https://dl.dropboxusercontent.com/u/71843785/Va...	Select Below	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Import CSV File					

Figure 2.

3 <<http://letters.nialloleary.ie/>>.

Two additional problems raised by our reliance on CSV imports were Excel's unfriendliness with long narrative entries (which many of our fields contain) and the imperfect fit between our fields and Dublin Core. Our solution for the former was simply to get used to it; the latter is discussed below.

The challenge of Dublin Core: element 'translation'

The CSV import option offered by Omeka maps our columns to the Dublin Core element set. Unfortunately for us, our metadata and the Dublin Core elements do not fit very well, nor does omeka.net offer the possibility of renaming the DC elements for public view, which means that we had to 'translate' them for the readers. For example, DC has no Addressee or Recipient element (a rather important field for a collection of letters), but we are using the Contributor element for this purpose. We did the same for Origin and Destination (absent in DC), which we mapped to the Coverage element in DC. In order to avoid confusion on the user end, we added 'Addressed to:' in front of recipient names and 'Origin' and 'Destination' in front of place names, as illustrated below. Surely not the most elegant solution, but it does help to clarify what we mean by 'Contributor' and 'Coverage'.

	Example from CSV File	Map To Element	Use HTML?	Tags?	Files?
Origin	Origin: Rome	Coverage 	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Destination	Destination:	Coverage 	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 3.

The challenge of controlled vocabularies: fuzzy data and spelling variations

Although we tried to have controlled vocabularies in as many columns as possible, we ended up with only five (Type, Language, Subject, Origin, and Destination). We also plan to standardize the Author and Addressee columns in the future, but for now they remain rather fuzzy. The rest of the columns contain complex information that cannot be satisfyingly conveyed with controlled vocabularies. One example is the 'Related' column. One letter can have several types of documents related to it (attachments, responses, decisions, comments and notes, etc.); moreover, these related items are not adjacent but scattered throughout our collection (our database is not growing chronologically or thematically but follows the organization of the archive). Therefore, as we add items, the relationships among them become more and more complex and our Related column keeps growing (Figure 4).

Spelling variations for obscure or uncertain names is a problem common to all document editions and we follow general usage on these matters. When no generally accepted spelling is known, we choose the one that occurs most often (in the case of personal names, the one used most often by its owner) and use it across

'Relation
<p>For the minutes of the PF meeting where this matter was discussed, see: APF ACTA 41, 305r, 330r-338r.</p> <p>With "the decree" Luigi Da Palermo probably means the decree issued by Di Seravezza on 1670-08-21: APF SOCG 430, 229r (database item 721).</p>

Figure 4.

the database, with alternate spellings within brackets. The standard and alternate spellings feature on our Biographical Dictionary page as well. We often have to make adjustments across the board as we discover new spelling variations.

User-experience limitations on omeka.net

Omeka.net is a fuss-free, low-maintenance platform with many advantages, but it certainly has its limitations. For one, its search function is rather confusing: it is far from user-friendly and it does not search attachments (although it claims it does). We found two partial solutions to this problem: we added search instructions to our user manual and we made our item summaries as detailed as possible in order to compensate for the non-searchability of the attached transcriptions.

The platform offers limited sorting options and interconnectivity. Links to other items can be added manually and we plan to do that in the future (in the Related field especially), but for now we are relying on tags, which can be sourced from any of our spreadsheet columns during a CSV import. Unfortunately, this procedure turns the contents of an entire cell into tags, which can look rather silly in the case of wordy entries (we ended up with tags such as 'Addressed to Giovanni Nicolo Guidi di Bagno', as seen in Figure 5). For this reason, we will no longer tag the Contributor (Addressee) element and we will reconfigure the Subject and Author ones in order to turn them into more meaningful tags.

Another limitation comes from Omeka's navigation options, which are not very flexible. We are still pondering how to make the distinctions between document types more easily apparent, as the explanations in our user manual are not very effective for this purpose. One solution would be to split the collection into sub-collections based on document type (in our case, the main distinction is between correspondence and minutes), but that would also fragment the search and tag functions. For now, all items are placed in the same collection.

BROWSE ITEMS

Browse All **Browse by Tag** Search Items Browse Map

Salé Madagascar Capuchins Christophe Authier **Petitions for faculties** Mission reports Alet Sedan Babylon Soissons Trabzon
Toul Ireland **Tunis Algiers** Varia Geneva CEC CM petitions CM statutes Galleys Cardinal rings St. Lazare **Letter** Philippe-
Emmanuel de Gondi Addressed to Bernardino Spada **Petition Addressed to PF** Addressed to Urban VIII Addressed to Guido
Bentivoglio d'Aragona **PF** Bernardino Spada **Minutes** Addressed to Giovanni Francesco Guidi di Bagno Giovanni Francesco Guidi di
Bagno Founding documents Delatie T. Gavot Addressed to Francesco Ingoli Vincent de Paul Louis Callon Antoine Portail Jean de la Salle Antoine
Lucas Joseph Brunet Jean Dehorgny Notes Francesco Ingoli Various SCER Father Hyacinth Addressed to Father Hyacinth Addressed to
Alessandro Bichi Statutes Contract CM Addressed to SCER Report Decree Claude Bouthillier Printed booklet Addressed to Univ. of Paris
Addressed to Ranuccio Scotti Henri Loys Addressed to Christophe Authier Girolamo Grimaldi-Cavalleroni Avignon Raymond de Lisle Trye Jean
Duval Nicolo Guidi di Bagno John O'Fahy G. Isvard Addressed to Nicolo Guidi di Bagno Addressed to Luigi Capponi Pierre de Piviers Assessor of
the Holy Office of the Inquisition Boniface Nouelly de Martin Louis XIV Anne of Austria Addressed to Henri Prat Jean Le Vacher Royal patent
Discalced Carmelites Charles Nacquart Addressed to Vincent de Paul Simon Legras Addressed to SCC Claude Dufour SCC Addressed to Nicolas
Pavillon Nicolas Pavillon Charles de Bourlon Addressed to [Card. Antonio Barberini Jr.] Addressed to [Luigi Capponi] Antonio da Genoa Giuseppe
Maria da Genoa Addressed to [Francesco Ingoli] Addressed to [Francesco Ingoli] 'Nicolo Guidi di Bagno de la Vigne and Maurin Pierre Gallais Yves
Vignal: Cordemoy: Corbin: Germani: Raqueneau: Mahaut Addressed to [Ottavio Bandini] Addressed to [Ludovico Ludovisi] [Bernardino Spada]
Origin: Rome Destination: Trinitarians Other secular missionaries Origin: Algiers
Destination: Rome Michel Monmasson Non-Catholic Christians Slaves (Christian) **Destination: [Rome] Origin:**
unknown Edme Jolly Tripoli **Destination: [Algiers] Slave priests [Eduardo Cibo] Other religious orders Summary The prefect of the Tripoli**
mission Destination: Tripoli Petitions for missionary patents Scotland Destination: Algiers Barbary (unspecified or several locations)
Conversion Destination: unknown [Urbano Cerri] Origin: Urbino Scottish Catholics abroad William Lesley Ransom Origin: Pesaro Renegades
Franciscans Archbishop of Armagh Jesuits Visitation [William Lesley] Venetian slaves in Algiers Tunis and Tripoli Origin: Tunis Fabrizio Spada
Francesco Gatta Destination: Naples Destination: [Bizerte] Destination: Malta Malta Slaves (Muslim) Francesco Gatta and Giovanni Battista de
Bonis The Scottish missionaries [Francesco Ravizza] Destination: [Paris] Origin: Bizerte Bizerte Marcello Costa Destination: Bizerte [Federico

Figure 5.

The advantages of using omeka.net

Despite its limitations, omeka.net has a number of important advantages. First of all, it is reasonably priced and it does not require intense technical assistance. Second, it offers flexibility, control, and the capacity to easily publish, modify, and update the outcome, all of which is much more difficult to do with more sophisticated tools that require the active involvement of digital specialists. (In the first phases of our project we often confused our DHO colleagues with our fuzzy data and constant changes!)

Most importantly, omeka.net offers enhanced educational opportunities that I had not foreseen initially. Because the platform is relatively easy to use and update, I can allow student assistants much more input than would be the case in a more controlled environment (they not only make transcriptions but also are engaged in creating and updating the metadata). In the process, they are confronted with in-depth source criticism: the constraints of Dublin Core, for instance, force us to cooperate in finding solutions for difficult questions about source authorship and other uncertain aspects of our documents. Student assistants also learn the importance of standardization as they advance in their familiarity with the database. Initially they would be relatively sloppy in their transcriptions and summaries, and my immediate corrections never helped as much as the subsequent tasks they received, which involved them in some aspects of our CSV imports or the creation of simple pages for our user manual. It was only when they had to use the database for those tasks that they truly realized how important precise spelling and standardized formats were. Allowing students to have this much input and

publishing our work as quickly as we do certainly has the downside of imperfect results – indeed, our database continues to contain errors and inconsistencies that are only gradually eliminated. However, this approach truly engages students in the process and teaches them the accountability of editing choices, which is a priceless learning experience.

Of general and homemade encoding problems

Daniela Schulz¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

The *Capitularia* project is a long-term venture concerned with the (hybrid-) edition of a special kind of early medieval legal texts, called *capitularies* because of their subdivision into chapters (lat. *capitula*). The textual tradition of these decrees is rather specific due to the ways the dissemination took place: mostly they were transmitted within collections carried out by attendants of assemblies where these texts were promulgated, or based on copies sent to bishops and other office bearers, creating quite a variety of different versions of the text that once had been presented.

The project is funded by the *North Rhine-Westphalian Academy of Sciences, Humanities and the Arts* since 2014 and will run for 16 years. It is being prepared in close collaboration with the *Monumenta Germaniae Historica* (MGH) and the *Cologne Center for eHumanities* (CCeH). A website (Figure 1) presents diplomatic transcriptions (in TEI-XML) of the various collections as well as manuscript descriptions and further resources. It is based on the WordPress-framework, and does not only serve for the presentation of the material to the public, but also as a central platform for exchange among staff, be it communication, data, resources or tools. On the website each text is presented as it appears in the respective manuscript. A critical edition, including commentary and translation into German will appear in print. The digital edition does not only provide the material for the printed edition, but is meant as its permanent counterpart with a focus on the tradition and transmission of the texts.

Capitularies cover a wide range of different topics (e.g. administration, religious instructions) and are specific to Frankish rulers. They seem to have originated as individual texts from deliberations and assemblies at court, but hardly any originals

¹ dschulz@uni-wuppertal.de.

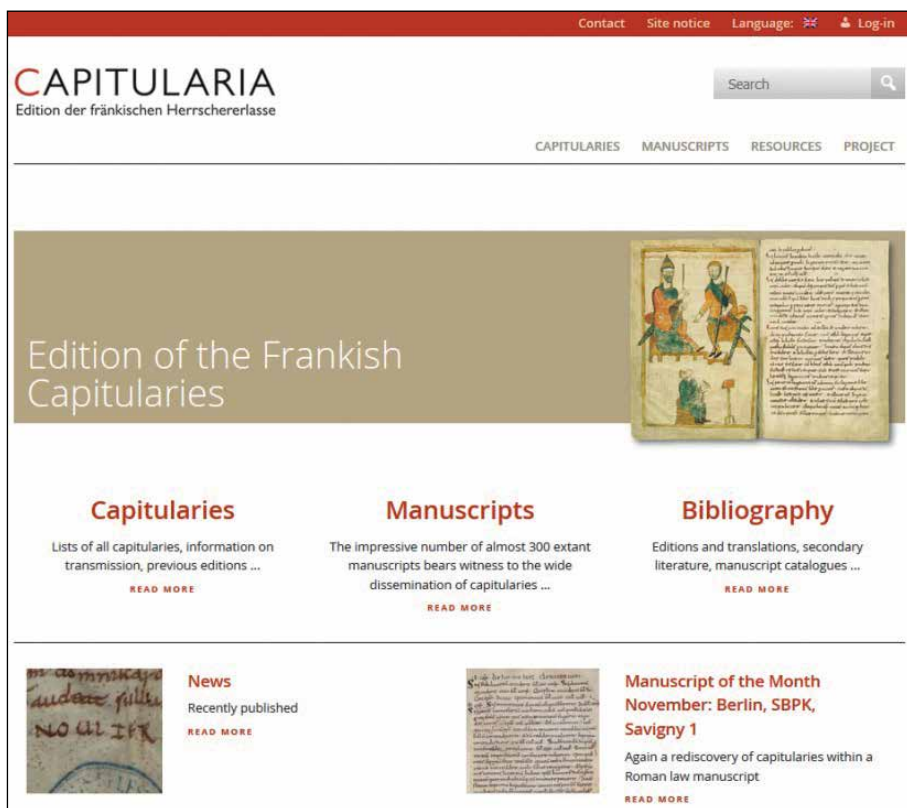


Figure 1: *Capitularia* Homepage (<http://capitularia.uni-koeln.de/en>).

have survived. What most capitularies have in common is their outer form as a list of chapters. Some appear to have been official documents; others might have been private notes, drafts or extracts. Different capitularies often are mixed into long lists of chapters. They were rearranged, modified, extracted by the (contemporary or later) compilers, which makes it hard to judge the status of a single manifestation of a text. These texts also differ significantly in length and number of witnesses. Given this great variety and heterogeneity, the modelling of an overarching structure to depict and reference the single textual units in their various manifestations within the manuscripts hence is one of the biggest difficulties. Overall there are about 300 texts in more than 300 extant manuscripts. The specificities of the source material pose particular challenges to the TEI encoding (regarding the identification of certain passages, overlaps, contamination etc.), and thus to the work of the editors. The project's long term adds to these complexities, since one hardly can anticipate future technological developments, which is why ensuring coherence is quite an issue. Today's decisions deeply determine future possibilities, and planning 16 years ahead does not seem achievable.

Since the start of the project, different kinds of encoding problems have arisen, resulting in constant revisions of the transcription guidelines. Besides overlaps in the textual structure as well as contaminations, the inclusion of tools such as computer-aided collation enforced further adjustments to the mark-up. Based

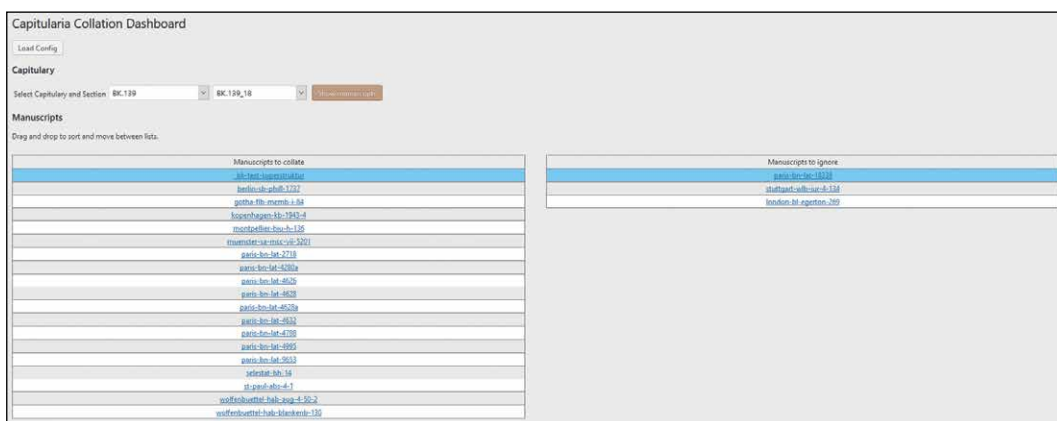


Figure 2: Capitularia Collation Dashboard; Manuscripts featuring Boretius-Krause No 139, c. 18.

bk-text-supplementstruktur	ussatorum	montemari	hps	facere	presumpterit	escagita	ictus	usupient	aut
berlin-dl-pub-1727	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
gotthe-fr-memb-184	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
lopenhagen-bk-1943-4	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
montpellier-bk-136	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
montserrat-ca-muc-vib-5201	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-2718	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4269a	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4628	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4628a	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4632	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4788	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-4995	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
parlo-bn-lat-5653	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
st-paul-aba-4-1	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
st-paul-aba-4-142	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
wolffenbuttel-hab-aug-4-50-2	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut
wolffenbuttel-hab-blankens-139	ussatorum	montemari	hps	facere	presumpterit	EX	ictus	usupient	aut

Figure 3: Collation output with the text of the Boretius-Krause edition as base text for comparison.

on *CollateX*, alignment tables are created to facilitate the editorial work with the numerous textual witnesses. This functionality was included in the *Capitularia Collation* plugin (Figure 2), which allows to compare any chapter of any capitulary in any set of manuscripts by either using the text of the old 19th century edition by Alfred Boretius and Victor Krause, or any version from a given manuscript as base to compare against (Figure 3).

The benefits of this procedure are quite obvious; however, the implications on the encoding still enforce a critical reflection and examination. Another encoding issue that will arise in the near future, and that will further add to the complexity of the mark-up is the question of how to identify, encode and then finally present certain collections of capitularies that seem to exist. Some capitularies apparently occur in conjunction with each other quite regularly, and some (non-official) collections already have been identified by previous scholars. To record and analyse these in a systematic way, a further ‘level’ of encoding needs to be inserted.

One further challenge still unresolved concerns the limits of creating a decent print output for a critical scholarly edition with all its components (text, translation, critical apparatus, and annotation) directly based on the XML-files. Due to the problems related to the application of *XSL:FO*, the critical print edition is prepared using the *Critical Text Editor (CTE)* as an (interim?) solution. The TEI-output offered by the CTE is fed back to the website to publish draft versions

(Figure 4). This output is hardly perfect, and in general a more integrated or interrelated workflow between digital and print edition is definitely desirable for the future.

In [*] nomine domini dei et salvatoris nostri Iesu Christi Hludowicus, divina ordinante providentia imperator augustus.	Im Namen des Herrn Gottes und unseres Heilands Jesus Christus, Ludwig durch die Anordnung der göttlichen Vorsehung Kaiser Augustus.
Quia [*] iuxta apostolum [*] , quamdiu in hoc saeculo sumus [*] , peregrinamur a domino	Weil wir gemäß dem Apostel, solange wir auf Erden sind, uns weiter von Gott entfernen und
<div> <p>Augustinus, <i>Enarrationes in psalmos</i> 36, 2, CC 38 S. 354: quamdiu enim hic uiuitur, crastinus dies semper ignoratur; <i>ders.</i>, <i>Sermo</i> 113A, Morin S. 152: negligentes non debemus esse: crastinus dies nescis qui sit. Vgl. auch D LdF 324, MGH DD Kar. 2, 2 S. 803 (833 für Saint-Denis): Quapropter, quia certus est quandoque venturus terminus et incertus formidatur eventus ...</p> </div>	<p>er gegenwärtigen Zeit fest, nichts ist, sondern alles in schnellem Lauf und weil wir nach dem Zeugnis der S, was wir können, sofort ausführen niemandem der nächste Tag</p>
operandum crastinus dies [*] promittitur, omnesque secundum [*] apostolum [*] ante tribunal Christi stabimus, ut unusquisque rationem pro his, quae gessit, reddat, nobis praecipue – qui [*] ceteris mortalibus conditione aequales [*] existimus et dignitate tantum regiminis [*] supereminemus [*] , qui non solum pro	<p>zugesichert wird, um Gutes auszuführen, und weil wir alle gemäß dem Apostel vor dem Gericht Christi stehen werden, damit ein jeder sich für das, was er getan hat, rechtfertigt, müssen wir uns insbesondere – die wir nach unserer Beschaffenheit den übrigen Sterblichen gleich sind und nur durch die Würde der Leitung überragen, die wir nicht nur</p>

Figure 4: Draft version of the 'Prooemium generale' = Capit. LdF No 5, olim Boretius-Krause No 137 (<http://capitularia.uni-koeln.de/en/resources/texts/ldf-bk137/>).

The role of the base manuscript in the collation of medieval texts¹

*Elena Spadini*²

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

This paper pursues procedures and algorithms for collating medieval texts. They are an exemplary case for demonstrating the importance of awareness and critical understanding in using computational tools, which is the wider frame of this research.

We perform textual collation for investigating the *varia lectio*, i.e. the variants. In the case of medieval texts, the results often are used for understanding the relationship among the witnesses. This is fundamental not only when following a stemmatic approach, but also for choosing a best manuscript (Bedier's *optimus* or the copy-text), which can be identified only through a comparative analysis. On the other side, understanding the relations among the witnesses may not be the goal when collating dated modern materials.

If the reasons why we collate are well known, the way we do it, especially when we do it manually, is less documented: handbooks and essays seem to take for granted this delicate task or summarize it in a couple of sentences.³ Direct and indirect experiences, shared with colleagues, suggest that manual collation generally implies the selection of a base witness and the record of the variants of all the others, which are read aloud by a collaborator or visually checked. A base manuscript for the collation is chosen according to completeness and absence (or limited amount) of *lectiones singulares*, but it is necessarily, at least partly, an arbitrary choice, as performed during a preliminary stage of the text's analysis. It is

1 I would like to thank Ronald H Dekker (Huygens ING) and Gioele Barabucci (University of Cologne) for having discussed with me the questions addressed in this paper.

2 elena.spadini@unil.ch.

3 The only exception to my knowledge being Bourgain-Viellard 2001.

not surprising, however, that the base manuscript for the collation often is retained as the base manuscript for the edition.

The selection of a base manuscript against which to compare all the witnesses is a common procedure for several reasons, all of them being practical ones: witnesses may not be available at the same time; comparing each of them with all the others would be highly time consuming; it is difficult to record, organize and visualize all the variants among a number of witnesses.

The role of the base manuscript has been questioned in the last decades, especially, but not only, by scholars using computational tools or preparing digital editions. Spencer and Howe (2004), whose research is grounded in textual criticism as well as in bioinformatics, uses a spatial metaphor to address the issue, which would sound familiar to textual scholars, as the relations among the witnesses are normally represented spatially, through a stemma, an unrooted tree, Eulero-Venn diagrams, textual flow diagrams or other forms of representations. They argue that ‘reconstructing a stemma is like reconstructing a map from the distances between points. (...) If we only know the distances between each city and London (equivalent to collating each witness against a base text), we cannot reconstruct the relative locations of the other cities’.

Another example of challenging the base witness role is the Parallel Segmentation method for encoding a critical apparatus in TEI, which allows for avoiding the use of a <lemma>, and for considering all the readings ‘as variants on one another’ (TEI Consortium 2016; *cf.* Robinson 2004). This is also the case for manual alignment of sections of the texts in a table or a spreadsheet.

These experiences indicate that collating with reference to a base witness is not the only possible choice; more importantly, they demonstrate that it is not the most appropriate way to proceed, when the aim of the collation is to understand the relationships among the witnesses.

Nevertheless, software for automatic collation did not explore the alternatives to the base text paradigm until very recently. In fact, all software for automatic collation implemented text alignment uses a base witness, except one. In this paper, we focus on two software packages that today are used widely and that both follow the Gothenburg model (*cf.* TEI Consortium 2011): Juxta and CollateX. The use of a base witness is documented clearly in each of the different outputs available in JuxtaCommons (<http://juxtacommons.org/>): in the Side-by-side view, only the differences between two witnesses are visualized; in all the other cases, the user must select a base text. CollateX, on the other hand, allows for collating without first selecting a base witness. These behaviours are due to the type of algorithm that informs the software.

Two kinds of algorithms are used for textual alignment: pairwise alignment and multiple alignment ones. In the first case, all the witnesses are compared to the base witness (step 1) and the results are merged (step 2), as in the example below:

Step 1.

A	Dalla	collina	si	vede	una	grande	casa	rossa
B	Dal	belvedere	si	vede	una	grande	casa	azzurra

Dalla) dal B
collina) belvedere B
rossa) azzurra B

A	Dalla	collina	si	vede	una	grande	casa	rossa
C	Dalla	collina	si	vede	una	piccola	casa	rossa

grande) piccola C

A	Dalla	collina	si	vede	una	grande	casa	rossa
D	Dal	belvedere	si	vedono	tante		case	

Dalla) dal D
collina) belvedere D
vede) vedono D
una) tante D
grande) om. D
casa) case D
rossa) om. D

Step 2.

Dalla) dal B, D
collina) belvedere B, D
vede) vedono D
una) tante D
grande) piccola C, om. D
casa) case D
rossa) azzurra B, om. D

A pairwise alignment algorithm shows the differences among a pair of witnesses: recalling Spencer and Howe's metaphor, it provides the distance between each city and the capital.

Multiple alignment, on the contrary, aims to calculate the variation among all the witnesses, giving enough information as to locate different cities on a map or to arrange manuscripts in a tree.

However, a multiple alignment algorithm is far more complex and computationally expensive, since its goal is to calculate the best alignment for the whole set of witnesses. As considering all the possibilities would just require too much time, different approaches and heuristics have been tested. The first attempt to apply multiple alignment to textual criticism makes use of progressive multiple alignment algorithms (Spencer and Howe 2004). This procedure consists of a first pairwise alignment between all the witnesses, providing distances upon which to build a guide tree; and then a second pairwise alignment is performed, where the witnesses are compared in order of similarity, as inferred by the tree. Aligning the most similar witnesses helps in finding correspondences and introducing white spaces, which is the strong point of this kind of algorithms. Some drawbacks are: calculating the first alignment is heavily computational expensive; establishing the

order for the second alignment is a NP-complete problem (commonly referred as the travelling salesman problem); the act of comparing these witnesses creates a super-witness that represents a serialization of the variant graph where the order of the tokens is set arbitrary, but may not be irrelevant.

An alternative to using a progressive multiple alignment algorithm is to use a non-progressive algorithm. However, at the moment their implementation in textual criticism is still experimental and no related documentation exists yet (*cf.* Sahraeian and Yoon 2013).

The only software for automatic collation – in its entire history – that uses a multiple alignment algorithm, partly progressive and partly non-progressive, is CollateX. Its results still prove to be problematic in certain cases. For instance, if we avoid using of a base witness, we consider all manuscripts to be on the same hierarchical level: this implies that the order in which the texts are fed as an input should not affect the program’s output. This is not always ensured by CollateX, nor of course by any of the other collation software packages, as shown in the example below. Since CollateX is the only collation software that explores a non-traditional path through the implementation of multiple alignment algorithms, some difficulties are in order. Exploring them allows a better understanding of the inner mechanisms of the software.

To conclude, multiple alignment algorithms are arduous and complicated: this is the reason why the great majority of the programs use pairwise alignment. However, even if making progress and developing these algorithms is a challenge endorsed by computer scientists, textual scholars can push research in this direction, can incite and foment developments, working together to create new tools that better correspond to their needs. Improving the existing software, creating new software, choosing which one to use (according to one’s needs), correctly reading the results and, last but not least, recognizing and making use of the innovative potential. All this is possible when the scholar opens the black box and investigates how the software works and why. In the end, raising awareness and critical understanding of using these tools may provide an answer to the popular question: is there a real change, in terms of methodology and results, in the use of computational tools?

A	B	C		B	C	A
the drought	the first march	the first march		the first march of	the first march of	the
of	of	of		drought	drought	drought
march	drought pierced	drought		pierced	-	of march
hath perced	-	hath perced		-	hath perced	hath perced
to the root	to the root	to the root		to the root	to the root	to the root
and	and	-		and	-	and
is	-	-		-	-	is
this	this	-		this	-	this
the right the drought of march hath				is the	-	the right the drought of march hath

References

- Bourgain, Pascale, and Françoise Viellard. 2001. *Conseils Pour L'édition Des Textes Médiévaux. École nationale des Chartes*. Droz.
- Dekker, Ronald H., and Gregor Middell. 2010. *CollateX*. <http://collatex.net/>.
- Dekker, Ronald H., Dirk van Hulle, Gregor Middell, Vincent Neyt and Joris van Zundert. 2015. 'Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project.' *Digital Scholarship in the Humanities* 30. 3: 452-470.
- Robinson, Peter. 2004. 'Rationale and Implementation of the Collation System'. In *The Miller's Tale on CD-ROM*. The Canterbury Tales Project. Leicester.
- Sahraeian, Sayed Mohammad Ebrahim, and Byung-Jun Yoon. 2013. PicXAA & PicXAA-R Web-Server. <http://www.ece.tamu.edu/~bjyoon/picxaa/>.
- Spencer, Matthew, and Christopher J. Howe. 2004. 'Collating Texts Using Progressive Multiple Alignment.' *Computers and the Humanities* 38. 3: 253-270.
- TEI Consortium. 2011. 'The 'Gothenburg model' A modular architecture for computer-aided collation. Last modified on 8 April 2011. http://wiki.tei-c.org/index.php/Textual_Variance.
- TEI Consortium. 2016. 'The Parallel Segmentation Method'. In *P5: Guidelines for Electronic Text Encoding and Interchange. 12. Critical Apparatus, Version 3.0.0*, Last modified in March 2016. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS>.
- Trachsler, Richard. 2005. 'Fatalement Mouvantes: Quelques Observations Sur Les Œuvres Dites 'cycliques.' In *Mouvances et Jointures. Du Manuscrit Au Texte Médiéval*, edited by Milena Michailova. Orléans, Paradigme.

A tailored approach to digitally access and prepare the 1740 Dutch *Resolutions of the States General*

*Tuomo Toljamo*¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Digital facsimiles are fundamentally data and therefore, in addition to being rendered to meet the eye, they also can afford algorithmic access to visual elements in imaged documents. This paper presents an exploratory case study where that affordance was used successfully: in the study, a highly tailored tool was developed to access cover-to-cover digitised images of a single document, to recognise select details in its layout structuring, and to use that information to prepare its contents automatically into a TEI-informed structured source. The document in question, belonging to a vast archival series, was the *Resolutions of the States General* for the year 1740. The aim of the study was to explore the computational use and usefulness of visual documentary evidence and to trial its exploitation with the select material. In this paper, the experiences from the study also are used to comment on tool building in the editorial setting.

The States General was the governing council of the Dutch Republic. The council met daily to decide on current political matters in such areas as foreign affairs, trade, defence, etc. (Nijenhuis 2007). These deliberations are recorded in the vast archival series of the *Resolutions of the States General* 1576-1796. The series, which was written down first by hand until 1703 and later set in print, and which spans unbroken for over 200 years and runs up to some 200,000 pages, has been recognised as 'a backbone source for the political history of the Republic' (Hoekstra and Nijenhuis 2012). But the breadth of the series has also posed a

¹ tuomo.t.toljamo@kcl.ac.uk.

formidable editorial challenge: in over a hundred years of editorial history,² the Huygens ING and its predecessors have yet to advance further than the year 1630. Currently, it is seen that continuing to edit the remaining unpublished period from 1631 to 1796 using traditional editorial methods is out of the question. Instead, and importantly to the present topic, the Huygens ING recently started several small-scale pilots to explore methods alternative to editing for the opening of the series (Sluijter *et al.* 2016).

Related to the above pilots, the case study under discussion was designed to explore the use and usefulness of visual information in the digital opening of a single document, the Resolutions for the year 1740. The motivating idea was that there seemed to be a wealth of information encoded in the layout of the highly structured document which perhaps could be harnessed usefully. The primary data consisted of a set of document images, which were available as a folder of 452 JPEG files. The document itself consists of two main parts: an index of about 105 pages that was omitted from processing because it was seen not to meet contemporary standards; and two series of resolutions which together run for about 783 pages.

The selected means of exploration was tool development. Applying a document-centric perspective, the development work was based on an extensive modelling of the document's logical and physical systems of structuring. In terms of logical structuring, the main body of the document is formed by a record of a chronological series of sessions corresponding to the council's daily meetings. Each of the sessions follows a specific structure: they start with a date and an attendance list; continue with a resumption summarising the previous meeting; and follow this by a listing of the resolutions made during the meeting (see Sluijter *et al.* 2016, 682). Importantly, this logical structuring is mirrored and supported by an intricate system of physical structuring that extends across topographical, textual and typographical levels. Its main features include: the text flows in columns, with two columns per page and four per spread; its layout makes use of regularised layout patterning, italicisation, centring and indenting; and its flow is demarcated with large initial capitals and vertical spacing, which are used to signal where new sections start and previous ones end. In addition, organising elements and reading facilitators, such as running headers, lines dividing columns, and catchwords, play a large part in the overall arrangement. Here, the models of these structuring systems were used to guide the tool development.

The developed tool³ embodies a method which leverages the structural models and works as follows: firstly, a selection of mundane elements of the physical structuring is recognised from the document images (e.g. running headers, column dividers, large initial capitals, vertical spacing, layout patterning of text lines);

2 The editorial efforts so far include a sequence of volumes of the so-called *Old Series* (1576-1609) and the *New Series* (1610-1625) published between 1915 and 1994, and an electronic edition (1626-1630) published in 2007; the editorial history is outlined in Hoekstra and Nijenhuis (2012) and Nijenhuis (2007).

3 The tool is written in Python using the Jupyter Notebook. It is developed as a processing pipeline which also integrates existing tools and resources, such as the Tesseract 3.05 OCR-engine and INL's Historical Lexicon of Dutch. Currently under active development, the tool will soon be released as open source.

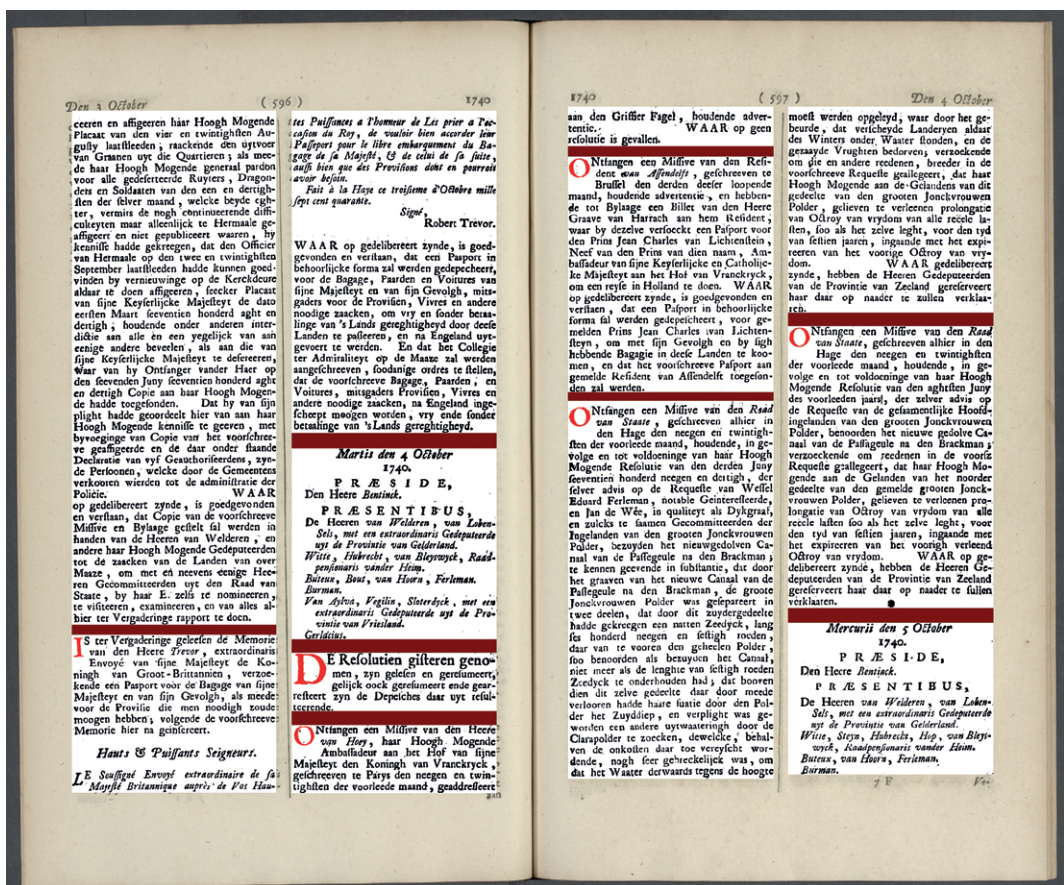


Figure 1: An example image of the document with an overlaid visualisation of the processing.

secondly, the selection's arrangement is compared to our model of the structuring in order to reconstruct the document's physical flow; and thirdly, the physical flow is used to infer the document's logical flow. In conjunction with the use of optical character recognition (OCR) to interpret the captured visual marks, the above analysis allows the tool to impose structure on the content and to store the information as an XML document. As a side note, the use of visual analysis also led to improved OCR results: because the logical roles of segments were known beforehand, it was possible to adjust the OCR engine's use of resources accordingly.

Although the layout elements and their useful patterns can be observed in the digital facsimile, they need to be computationally accessed through image data. For this, the work borrowed a set of simple concepts and methods from the field of document image analysis and understanding; these included, for example, binarisation, projection profiles, and connected components (for an introduction, see Doermann and Tombre 2014). Although simple, they are conceptually very powerful in allowing one to see image data and how to work with it in a new light.

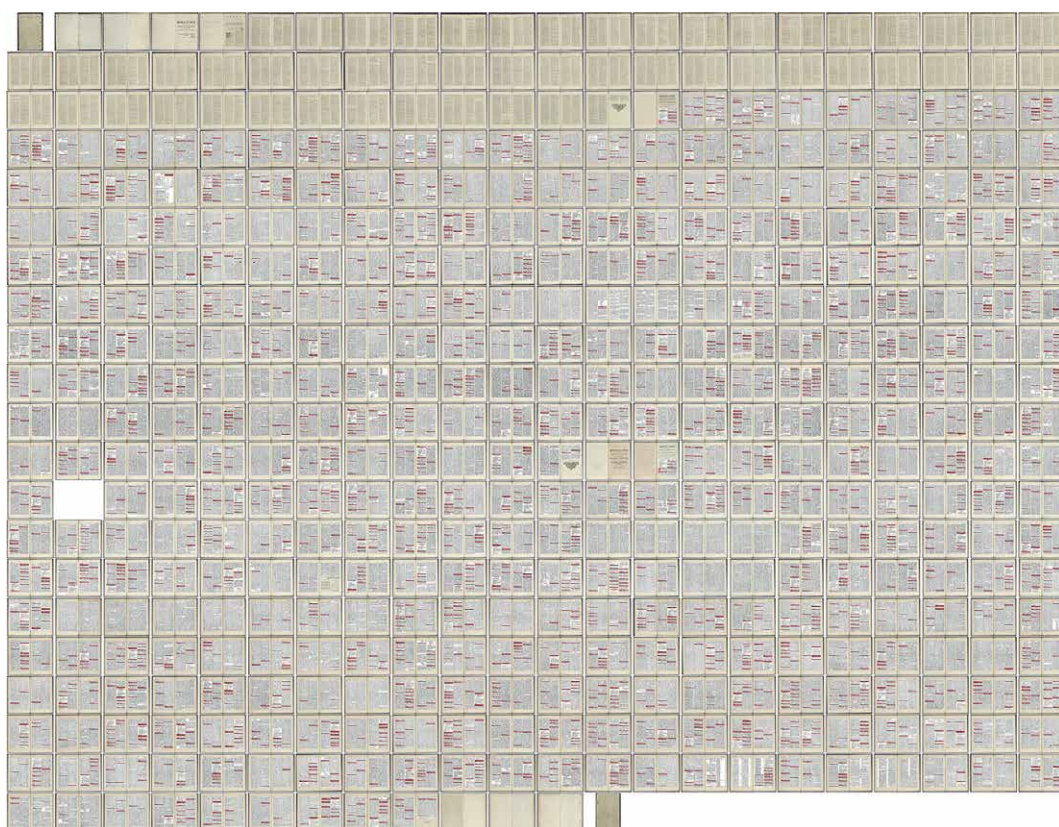


Figure 2: Digital facsimile of the whole document with an overlaid visualisation of the processing.

In this case study, the described method for exploiting visual information proved useful. The developed tool was able to help in the digital opening of the document by accessing the content and its structuring through digital facsimiles and by storing the information as a TEI-informed XML source. Because the series' manual editing is currently out of the question, the produced source while imperfect is still useful as an additional layer of access (Hoekstra and Nijenhuis 2012) and as a starting point for further manual or automated enrichment (e.g. by applying the methods explored in the other pilots).

This paper concludes by using the gained experiences to comment on tool building in the editorial setting. Firstly, the work suggests that more attention should be paid on tool development incentives, and that expectations towards the tools should be adjusted accordingly: for example, here the tool was built as a process to learn and the outcome could be described at best as research software, and at worst as something close to PhDWare.⁴ Secondly, the work highlights

4 'PhDWare' is a rarely used but fitting term: it describes research software written by people who often lack formal training in programming; whose main incentive is instead to write e.g. a PhD or a publication; and who are ill-incentivised to spend excessive amounts of time refining their code. It implies software that probably has worked at some point, but is often hard to build on or integrate, poorly documented, and likely not maintained.

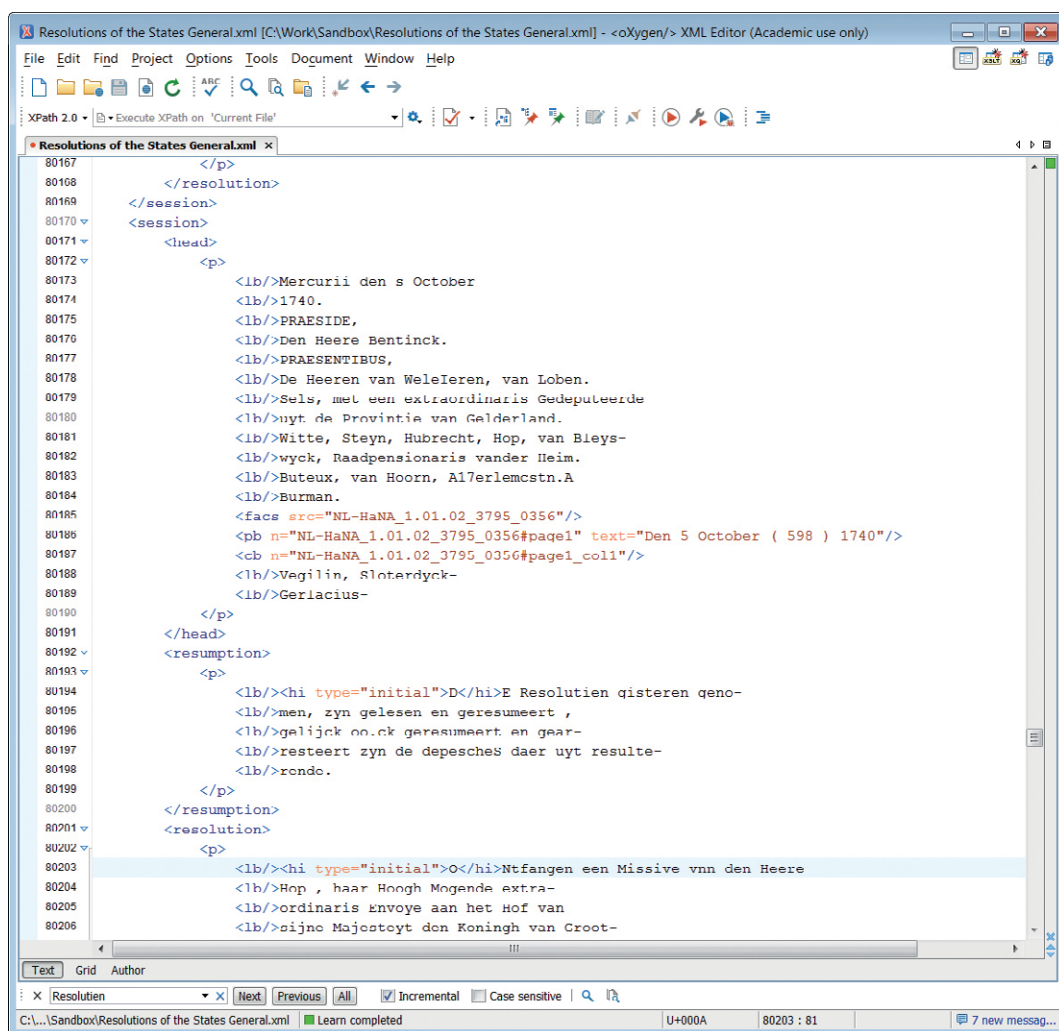


Figure 3: A screen capture of the generated XML source.

differences between out-of-the-box and customised tools: in terms of OCR performance, for example, customisation here allowed to turn the intricate layout patterning from a potential hindrance into a great enabler. Customisation also was found to invite more customisation, because the steps already taken bring into sight new low-hanging fruit. In the editorial setting, this seems to suggest that although editors cannot always be involved in driving technology, they should very much be involved in its informed application. Thirdly, the work exemplified that although the developed tool is highly tailored and as such not directly applicable to other material sets, it can at the same time be quite re-usable on both conceptual and code levels. Overall, the development process also pointed towards the importance of grassroots aggregations of knowledge, *e.g.* workshops and how-to documents focussing on specific editorial activities and on the current approaches to supporting them, in helping to gradually improve the current situation with editorial tools.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° 317436. The work was carried out during a DiXiT research secondment at the Huygens ING (KNAW).

References

- Doermann, David Scott and Karl Tombre (eds). 2014. *Handbook of Document Image Processing and Recognition*. Springer.
- Hoekstra, Rik, and Ida Nijenhuis. 2012. 'Enhanced Access for a Paper World.' *Paper presented at the Ninth Annual Conference of the European Society for Textual Scholarship*, Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam. Accessed 23rd March 2016. https://pure.knaw.nl/portal/files/1672164/Enhanced_Access_for_a_Paper_World.pdf.
- Nijenhuis, Ida. 2007. 'Resolutions of the States General 1626-1630: Introduction.' Accessed 18th November 2016. <http://resources.huygens.knaw.nl/besluitenstatengeneraal1576-1630/BesluitenStaten-generaal1626-1651/en/inleiding>.
- Sluijter, Ronald, Marielle Scherer, Sebastiaan Derks, Ida Nijenhuis, Walter Ravenek and Rik Hoekstra. 2016. 'From Handwritten Text to Structured Data: Alternatives to Editing Large Archival Series.' Paper presented at the Digital Humanities 2016, Kraków.

Editorial tools and their development as a mode of mediated interaction

Tuomo Toljamo¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Encouraged by the Convention's exploratory aims, this paper adopted two viewpoints: firstly, to view the production and sharing of editorial tools as mediated communication; and relatedly, to view a user's engagement with a tool as material interaction. It did so in order to pose the following question: Is there something to be gained by thinking along the above lines, for example, with regard to how editorial tools are produced?

The above question was motivated by the many issues recently voiced regarding editorial tools and the means suggested for addressing them. For example, tools often are regarded as black-boxes or as reinventions of the wheel. Their development efforts are reportedly hampered by language problems and differences in thinking between scholars and developers. Additionally, such statements as 'tools are built, but no one uses them', 'we need more tools, better ones' and 'instead of managing it, tools often add complexity' are not uncommon for one to hear. Complaints also have been voiced over how there is too much custom software which does not build on existing work, and how the resulting tools rarely are maintained properly.

Many such issues have been addressed in literature and within collaborative settings. For example, improvements have been proposed to various technical matters and production processes (e.g. relating to debates contrasting larger-scale research environments vs. smaller tools; see *e.g.* van Zundert 2012, van Zundert and Boot 2011). These especially have considered the use of resources with regard to long-term returns. Additionally, initiatives on various levels have sought to bring people together to make things, to build communities around tool development

¹ tuomo.t.toljamo@kcl.ac.uk.

(e.g. Interedition). Editors also often have been advised to standardise their work and to learn more about technology.

From this paper's perspective, however, it could be argued that an over-emphasis on technicalities and features can act as a distraction from other valuable viewpoints. One such viewpoint could be that of communication. In that context, the so-called Wii's Laws (see Korpela 2010), humorous observations underlining the difficulty of communication (e.g. 'communication usually fails, except by accident'), remind one of the involved challenges. Indeed, alongside such considerations some of the tool issues seem to present themselves in a different light. Thinking in these terms also might allow one to identify new issues or to propose different solutions for the existing ones.

In terms of mediated communication, this talk considered the production and sharing of editorial tools as communication mediated by the tools themselves. That is, while tools are developed for specific purposes, they are often also complex products used to communicate, to interact, to collaborate with others. Along these lines, if we would like to improve how we communicate by producing and sharing tools, it would be beneficial to understand how users interact with them.

For examining that interaction, this paper borrowed a framework of material interaction from Dant (2005, 2008). This informed the examination of how users interact with tools to consider: how intentionality is designed into objects; how users try to 'read' objects to understand them, and how they do so especially in accordance to their undertakings; and how this reading is done by means of perception, gesture and recovery – i.e. observing the object, manipulating it, and trying to recover from the unexpected by working out what happened and why. Additionally, the framework shows the production and sharing of tools as a mass medium whereas notably the users' interactions with them are particular and singular. Importantly, then, the user's reading of a tool is and can be situated consciously: it is affected by background and experiences, but can be primed with, for example, instructions and documentation for helping the user to read the tool in the intended manner.

Editorial tools are often complex objects. They implement conceptual models (for modelling, see McCarty 2005). In simple terms, these models embody selections of what is important and how it should be dealt with. Here, examined from the point of view of communication, it becomes all the more clearer to see how editorial tools are rooted in editorial theories and concepts, and how they also are developed in order to instantiate and disseminate particular theoretical and methodological positions. While this suggests that editorial tools embody the needs of the scholarship they are designed to serve and that they need to evolve alongside improvements in that understanding, it also seems to suggest that, in terms of material interaction, the tools often present themselves as icebergs.

Considered as icebergs, it seems that the users can only hope to be able to read what has been made available on the surface of these complex objects because it is challenging to correctly infer what lies beneath. That is, to infer those more implicit aspects of what the tool is about and why it is as it is. At the same time, it seems that an understanding of those aspects often would assist in the correct reading of the surface. For example, while a tool might seem like a reinvention of

the wheel, the reasons for its existence might well be hidden in the model below and only manifest themselves in less straightforward ways. This seems to suggest that it is important to situate tools carefully, also in other than technical terms. Indeed, for editors this is a familiar line of thought: introductory essays explain the underpinnings of editions and set the context for their reading, for expectations.

In general, the adopted viewpoints also seemed to emphasise how small, simpler tools communicate and are communicated with more easily and how black-boxes violate Dant's reading mechanism of perceive, gesture, recovery. They also seemed to emphasise the importance of investing deeper collaboration in tool development. There, the discussed considerations could feed back to how tools are developed, and especially to how their reading is situated, in order to better reach common aims of development efforts. For example, these kinds of questions also could be asked in specifying a new tool – in addition to laying out use cases, features and technicalities.

In conclusion, we seem to be at a risk of the following pattern: we build tools and share them; others use them, read them, and misunderstand them. However, following Wiio's Laws, by paying attention to aspects of communication we perhaps can hope to increase the chances for accidental successes.

References

- Dant, Tim. 2005. *Materiality and Society*. McGraw-Hill Education (UK).
- . 2008. 'The 'Pragmatics' of Material Interaction' *Journal of Consumer Culture* 8, no. 1: 11-33.
- Korpela, Jukka. 2010. 'A Commentary of Wiio's Laws', last accessed 24 August 2015.<<http://www.cs.tut.fi/~jkorpela/wiio.html>>.
- McCarty, Willard, 2005. *Humanities Computing*. Palgrave Macmillan.
- Van Zundert, Joris. 2012. 'If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities' *Historical Social Research / Historische Sozialforschung*, 165-186.
- Van Zundert, Joris, and Peter Boot. 2011. 'The Digital Edition 2. 0 and the Digital Library: Services, Not Resources' *Digitale Edition Und Forschungsbibliothek (Bibliothek Und Wissenschaft)*, 44: 141-152.

TEI Simple Processing Model

Abstraction layer for XML processing

Magdalena Turska¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

The Guidelines of the Text Encoding Initiative Consortium (TEI) have been used throughout numerous disciplines producing huge numbers of TEI collections. These digital texts most often are transformed for display as websites and camera-ready copies. While the TEI Consortium provides XSLT stylesheets for transformation to and from many formats there is little standardization and no prescriptive approach across projects towards processing TEI documents.

A Mellon-funded collaboration between the TEI Consortium, Northwestern University, the University of Nebraska at Lincoln, and the University of Oxford, TEI Simple project aims to close that gap with its Processing Model (PM), providing the baseline rules of processing TEI into various publication formats, while offering the possibility of building customized processing models within TEI Simple infrastructure. For the first time in the history of TEI there exists a sound recommendation for default processing scheme, which should significantly lower the barriers for entry-level TEI users and enable better integration with editing and publication tools.

Possibly of even greater significance is the layer of abstraction provided by TEI PM to separate high-level editorial decisions about processing from low-level output format specific intricacies and final rendition choices. PM aims to offer maximum expressivity to the editor, at the same time encapsulating the implementation details in TEI Simple Function library. A limited fluency in XPath and CSS should be enough to tailor the default model to specific user's needs in a majority of cases, significantly reducing time, cost and required level of technical expertise necessary for TEI projects.

¹ magdalena@exist-solutions.com.

We believe editors indeed can become more independent from developers in tweaking the processing rules especially in numerous cases where editors already are deeply involved in the decisions about encoding on the level of XML markup and do have some rudimentary coding skills. Preliminary results show that it is perfectly reasonable to expect editors to tailor the existing high-level processing models to fit their specific needs, especially if lightly supported in concrete XPath/CSS formulations. For the non-technical users in particular the effect of incorporating the Processing Model into eXist-db native database and application framework environment makes it a viable option for out-of-the box publication of TEI documents, tremendously softening the learning curve necessary to achieve the same effect otherwise.

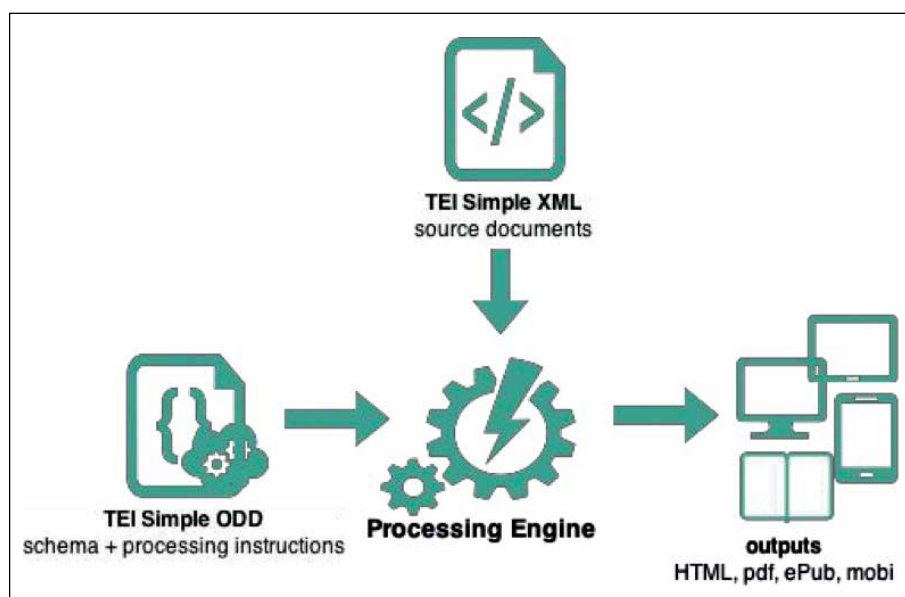


Figure 1: TEI Simple Processing Model.

Schema proposed by TEI Simple project assumes that TEI ODD files enhanced with processing `<model>`s directives are processed together with TEI XML source files to arrive at output document in one of supported formats. At present the most mature PM implementation is the eXist-db one written in XQuery². In this scenario editor needs to install TEI Publisher (TEI Simple Processing Model app), available from eXist-db's Package Manager, upload TEI source files and, if necessary, customize default TEI Simple ODD file while the rest of the process is taken care of by the application.

2 Wolfgang Meier, TEI Simple Processing Model <http://showcases.existdb.org/exist/apps/teisimple/index.html>

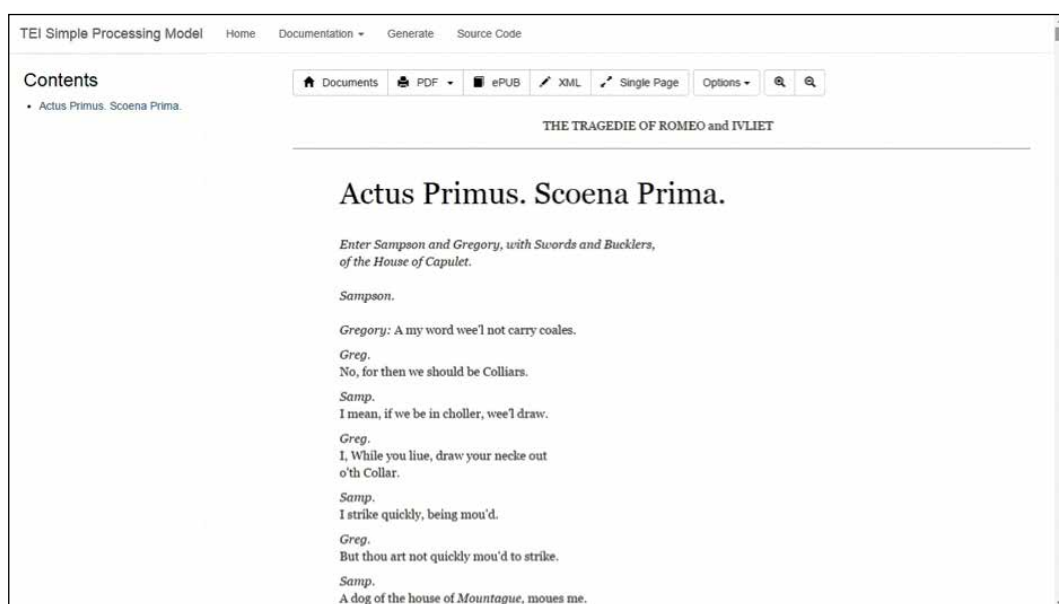


Figure 2: Default TEI Simple Processing Model output for Romeo & Juliet.

Most important extensions to TEI ODD metalanguage concentrate on `<model>` element and its children that can be added to existing TEI `<elementSpec>`s.

```
<elementSpec mode="change" ident="abbr">
  <model behaviour="inline">
    <outputRendition>text-decoration: underline;
  </outputRendition>
</model>
</elementSpec>
```

Figure 3: `<model>` example: abbreviations.

Models need to have at least the `@behaviour` attribute specifying function from the TEI Simple function library to apply. Parameters, where applicable, can be passed via `<param>` children of a model. In cases when different treatment is necessary depending on element context (e.g. to distinguish between headings for divisions of type='act' and others) all possible situations need to have separate models identified via XPath expressions on `@predicate` attribute. Furthermore general rendition directives can be recorded as CSS instruction in `<outputRendition>`. More in-depth discussion of encoding scheme and full documentation can be found at TEI Simple GitHub page.³

³ Sebastian Rahtz et al, TEI Simple, <http://teic.github.io/TEISimple/>

The processing model is a new proposal and needs to be tested extensively before announcing it a success, nevertheless it is employed already in production by real world projects, some of which have been running for a significant number of years and have already produced vast collections of material as exemplified by historical documents of the US Department of State.⁴ The Office of the Historian publishes a large collection of archival documents of state, especially those appertaining to foreign relations. Having the material previously published with custom-built XQuery/XSLT packages means that it is possible to compare the results of using an approach based on the processing model with the previous one in terms of the quality and consistency of final presentation but also in more quantitative ways like the size of necessary code base, development time and ease of the long-term maintenance.

The first challenge in such an endeavour, obviously, is rephrasing the transformations previously formulated in XQuery/XSLT using ODD meta-language extensions proposed by TEI PM. Preliminary results are very encouraging even though, as expected, it became necessary to extend the behaviours library to accommodate some specific needs. From the developer's perspective it is immediately clear that using the TEI processing model brings substantial advantages in development time and results in much leaner and cleaner code to maintain. For the Office of Historian project figures suggest code reduction by at least two-thirds in size, from original 1500 lines to mere 500. Numbers are even more impressive realizing that the resulting ODD file is not only smaller, but much less dense code, consisting mostly of formulaic<model>expressions that make it easier to read, understand and maintain, even by less skilled developers.

TEI Processing Model does not exist in a void, to the contrary, it was conceived to be combined with existing XML technologies to arrive at promising technology stack for creation, publication and reuse of scholarly resources. It is hoped that editors, curators and archivists as well as developers dealing with TEI will benefit from employing TEI PM in their workflows.

⁴ Foreign Relations of United States, <http://history.state.gov/>

WP3

Academia,
Cultural Heritage, Society

Edvard Munch's writings

Experiences from digitising the museum

Hilde Bøe¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Introduction

In this paper I will delve into the details of a few of the experiences, challenges, plans and hopes that you are sure to come across during a project like the digital archive of Munch's writings. In the first half I will be talking about experiences and challenges focusing on manuscripts; the written languages, the handwriting styles and variants as well as the many senders and receivers. In the second half I will discuss our plans to finish the still on-going work and to reflect upon what lies ahead and our ambitions for a future Munch web. But first a little bit about the collection and the digital archive.

The Munch museum's collection stems from a large part from Edvard Munch as he bequeathed his belongings to the City of Oslo in his testament. Among these was an archive of writings. Munch seems to have collected correspondence and writings almost regardless of the content. Today we have approximately 25,000 manuscript pages in the museum, most of them from Munch's own home. Over the last eight years the Munch Museum has been working with its digital archive on Edvard Munch's writings and correspondence. The digital archive was launched in 2011, but the work is still on-going. 3/5 of Munch's own writings have been published as well as half of the correspondence addressed to Munch, and a selection of about 1100 pages of his notes on art and his literary sketches have been translated in to English and published at emunch.no.

Munch's writings have been a somewhat hidden treasure and were little known. They have been registered as museum objects, i.e. with seemingly little thought of

¹ hilde.boe@munchmuseet.no.

what they are or what belongs together from a textual perspective, and although they have been transcribed, not much research has been done. Being a private archive of notes (of 'all kinds'), literary sketches, prose poems, letters and letter drafts they are often hard to categorise precisely as they lack proper genre features, dates etc. Aside from the sent letters, most of the texts are in an unfinished state of some kind and there are also often several versions.

Much better known of course are Munch's artworks, be it his paintings or the graphical prints. The collection also comprises some 7000 drawings as well as furniture, family letters, artist equipment (1000 paint tubes!) to mention some. In the coming years the museum will be working on creating an online presentation of all objects in its collection. Many of them are related to each other since Munch often worked on the same motifs in whatever media he chose.

The Munch archive has been built on the scholarly principles of modern philology (*Neue Philologie*). It is not a typical scholarly edition, rather a source edition. It has scholarly transcriptions of original sources, and offers these in a synoptic view that combines the manuscript facsimile and the diplomatic transcription. It is thus a hybrid between the facsimile edition and the diplomatic edition. The Munch archive can also be viewed as an analogue to the *Catalogue Raisonné*, i.e. the complete catalogue of an artist's works within a genre, describing the objects without going into thematic analysis. The archive thus combines traditions from scholarly editing, museum cataloguing and art history.

Experiences and challenges

Handwritten manuscripts in several languages

Munch's own writings are in Norwegian, German and French while the correspondence for Munch has letters in Norwegian, Swedish, Danish, German, French and English as well as the odd occurrence of Italian, Spanish, Polish and Czech. We work almost entirely with handwritten material, and alongside Munch's own handwriting we have several hundred handwritings to consider when we transcribe and proofread manuscripts. The texts span 70 years, from 1874 to 1944. In addition to the numerous variants of personal handwriting styles we also have the changing handwriting script types of Cursive and Gothic with the German variants of *Kurrentschrift* (including *Sütterlin*) which of course sometime overlap in personal handwritings where the writer learned writing one script type and later learned a different type.

Transcription and Munch's written languages

To read handwritten texts is demanding, at times extremely difficult, just as Munch's handwriting can be. Handwriting distinguishes itself from printed type by being *imprecise*. Handwriting is affected by historical conditions, individual qualifications and by motivational factors in the writer's situation. It is affected by the physical and mental circumstances within and around the person writing, by the tools of writing and by the surface that is being written on. Handwriting will thus always be open to a degree of interpretation. It also means that one will

find places in the text where one must choose between several ways of reading, where the text remains uncertain; one will find places in the text where one cannot decipher the word or letter that has been written, resulting in holes in the text.

Combine this with Munch's faulty or lacking foreign language knowledge and – more often than not – sloppy practice, reading and transcribing Munch's texts becomes quite the challenge. To be able to read and understand (and then transcribe) Munch's German or French texts does not only require thorough knowledge of German or French, but also of Norwegian as it was written in Munch's time, as many of the errors and peculiarities of Munch's texts in German and French can be explained only as consequences of the writer's Norwegian mother tongue. *E.g.* knowledge of sentence patterns in Norwegian and of how a French word is pronounced (by a Norwegian) and then spelled 'orthophonically' by a Norwegian is if not necessary at least very helpful when reading Munch's French.

I think it is easy to take for granted the underlying need to understand the language of a handwritten text, and therefore to forget or underestimate the role this has on understanding the content of the text and on being able to transcribe it, and in the end the impact on the quality of the transcription. Transcribing depends first on understanding the language and then the content of the text, so if the text is written in 'poor' language with untidy or sloppy handwriting, the task becomes all the harder. If the transcriber is not very familiar with the language, either because it is foreign to the transcriber or an older (and thus unfamiliar) version of the transcriber's mother tongue, the transcriber's job will be very demanding, not to say at times impossible.

I am going to give you an example from Munch's Norwegian texts of one decision we had to make. Passing into adulthood Munch changes his handwriting and – among other things – almost always writes the *i* without the dot. The lack of a dot over the letter *i* can cause problems since the letter without the dot often is formed in the same way as *e* and can thus be confused with *e*. In some cases this leads to a new meaningful word, although luckily not always a word that is meaningful in the context. Or it could lead to modern versions of the same word, as in the case of the Norwegian versions of the words you, me, oneself, where *dig*, *mig*, *sig* with an *i* are transformed to *deg*, *meg*, *seg* with an *e*. We know Munch wrote *dig*, *mig* and *sig* when he was young because he dotted the *i*, and perhaps as an adult and an old man as well. I say perhaps, because we cannot know this for sure when he no longer dots the *i*. With the orthographic reform of 1938 *dig*, *mig* and *sig* were replaced with *deg*, *meg* and *seg*, and by then many progressive language users had used *deg*, *meg* and *seg* for years already. We have *chosen* always to transcribe the words as *dig*, *mig* and *sig*. There are other cases where both *i* and *e* create a meaningful word, and were we have also used our knowledge of Munch's writing habits, the history of written Norwegian and informed judgement to determine whether the word should be written with an *i* or an *e*.

The many senders and receivers

The manuscripts often refer to places, persons and institutions as well as artworks, books and other cultural objects. A commentary therefore is required, but since our resources have been limited we have had to prioritise. We are publishing

comprehensive person and institution registers and a commentary that is not at all comprehensive. We have chosen the briefest possible format on the person and institution descriptions focusing on the relation to Munch and where possible linking to other resources for fuller biographies. But, many of the persons in our material are hard to identify. Munch kept not only letters, but receipts, electricity and phone bills, and therefore all kinds of persons and institutions are mentioned and are senders and receivers. Secretaries and officials signing correspondence from public institutions are not included in our register, but still many remain that are hard to identify.

Related to persons are also the issues of copyright and privacy. It is more than 70 years since Munch passed away, but many of the people he corresponded with are not devoid of copyright yet. The copyright on some letters does not expire until 2078! For some of the senders we might contact heirs to obtain permission to publish the letters, but as this is also a really time- and resource-consuming task – all heirs must be found, asked and agree – we will consider the importance of the letters in a Munch context before doing so. The privacy of people *mentioned* in the letters also needs to be considered. Letters where living persons are mentioned shouldn't be published at all. In cooperation with the juridical department of the City of Oslo, we have drawn guidelines stating that letters should not be published until the persons discussed or mentioned have been dead for 25 years. Privacy concerns also guard the content of letters. If letters contain discussion of sensitive matters, *e.g.* disease conditions, criminal offenses, sexual relationships, but also ethnic origin, political, philosophical or religious convictions, we ought to be careful and perhaps refrain from publishing. Although we read every letter during their preparation for publishing, having 10,000 letters and letter drafts to consider on copyright and privacy matters makes this task formidable, and we therefore have decided that we publish 'in good faith' where we have not been able to identify a person's year of death or where we are uncertain about the sensitivity of the letter content. If heirs later come forward with protests, we of course can remove letters should it be deemed necessary.

Plans and hopes

Finishing the work

As I mentioned at the beginning of my talk, we have not yet finished the work, neither with Munch's texts nor with his correspondence. The reasons are several, but the aspects I outlined above are the main ones: This material is *so* complex, it takes a *lot* of time building scholarly reliable, digital and authentic representations of it. There are also other tasks left on our to-do-list, and these are complex, resource intensive tasks too, and I hope that we can use digital tools to aid us with these tasks:

- Analysing and visualising the geographical encoding we have painstakingly added to our XML files using <placeName>tagging is high on my list. Visualising through maps is such a good way to convey the scale of networks and travelling, the extent of Munch's impact etc.

- We have not had time to connect letters in correspondences. We offer lists of letters to and from a certain person, but would like to connect the letters in the order of the original correspondence between sender and recipient
- We would also like to link Munch's drafts for a letter to the sent letter or at least to each other. It might not be possible to arrange them in chronological or genetic order, but we can say that they are versions of each other
- There are also other types of relations and connections between texts in our collection that I would like to make explicit. There are *e.g.* 'broken' texts where parts for some reason have not been catalogued as an entity or Munch for some reason has finished writing a text in a different part of the book than where he started out. Prose poems and literary sketches also come in several versions, which definitely should be linked to each other.

A Munch web

The texts do not exist alone though; they are connected to other objects in our collection. In 2019 the Munch museum is moving to a new building and ahead of the move we need to go through every object in our collection to make sure it is registered properly and that its condition allows moving. For this we receive extra funding and my hope is that there will be funding to finish digitising the texts as well. This hope is connected also to the fact that we are planning to build a digital collection presentation, an eMuseum, which will be launched when we move into our new building. We think it is a good opportunity to do this as we review our collection for moving. That process will give us a complete overview and will allow us to bring the data standard of all records in the database up to a minimum level. Everything will also be photographed, making it ready for sharing. Our collection is not very large. We have some 45.000 objects altogether. We are a single artist museum where all objects are related to Edvard Munch in some way, and they are also – more often than not – connected to each other. It is the connections between all our objects that are going to be the basis for the Munch web that will appear when we have built our eMuseum.

From my perspective the future of the online Edvard Munch collection is connected also to Linked Open Data (LOD). To quote europeana.eu's definition: 'Linked Open Data is a way of publishing structured data that allows metadata to be connected and enriched, so that different representations of the same content can be found, and links made between related resources.' As a museum I believe it is important for us to share our data when we publish them online. We will present the collection and the metadata we will release as LOD will be drawn from the collection presentation data, so releasing the data as LOD does not add much to the workload. Without the LOD version the collection presentation is much less worth – at least thinking towards the future and the semantic web. This is of course because releasing LOD makes it possible to connect to others and let others connect to us, adding to the online knowledge base.

We must of course enrich our own data using our expertise and our knowledge of what is in our collection. But we need to remember that we do not know everything – when we share data openly other data can enrich them further from

other perspectives. New knowledge will come from looking into what links reveal about forgotten or previously unseen relations. As a basis for learning and studies LOD is the future, and for visitors the past will come even more alive through the connections that are opened by LOD.

Working with emunch.no has taught me that the Web is home to so much information that just sits there waiting to be intertwined. I would like to show you an example of what I want to see happen. In his so-called 'Violet journal' Edvard Munch writes from France (and I quote from the English translation): 'Yes it is this nasty cough that will not subside – I should perhaps take some Géraudel pastilles.' In the original Norwegian text the two words that translate as 'Géraudel pastilles', are hard to make sense of, as Munch's orthography is faulty and his handwriting a challenge to read. After consideration and discussion including searching the web, we decided that he probably meant the French throat lozenges Pastille Géraudel. The Pastille Géraudel is in fact still quite famous and remembered even today due to the advertising campaigns that were really successful in large part because of the talented illustrators that were hired. These particular posters were created by Jules Chéret, a French painter, draftsman, printmaker and poster artist. Original posters and reproductions of them are sold to this day, you can even buy phone covers illustrated with a poster for the Pastille Géraudel, and of course they also appear in museum collections.

The throat lozenges obviously are not very important in Edvard Munch's text, they are mentioned in passing, and they are even less important in the Munch collection (but I will add in parenthesis that Munch suffered frequently from colds and bronchitis so he might have become very well acquainted with the pastilles during one of his stays in France). Still, online their link to a much wider cultural history can open up the text to the world and teach the reader of the 'Violet journal' something new and fascinating, and the other way around, they could bring people browsing catalogues for French illustrators to Edvard Munch – if the linking possibilities are there. How do we facilitate such linking? We use *e.g.* ULAN of the Getty Vocabularies to reference artists that illustrated the campaigns of the Pastille Géraudel. In fact, looking up Jules Chéret in the ULAN reveals an indirect connection between him and Munch as Chéret's student Eugène Carrière was the teacher of Munch's friend Norwegian artist Ludvig Karsten.

In conclusion

To conclude, digitising the museum is a large, complex and resource intensive task on many levels even if the collection in itself is not that large. It is necessary to spend the required resources to produce data of high quality. With high quality data – whether these are the digital images, the transcriptions or the metadata – we have the best base for whatever we want to build or do. This might seem obvious, but with the everyday battle for limited resources that most of us in cultural heritage institutions fight, and faced with the expectation of superiors regarding not only high quality, but also immediate (or at least *faster*) results it is not easy to convince said superiors that the slow, steady pace you are following actually is *needed* if the desired quality is to be achieved. This is one of those occasions where *to hasten slowly* really is the best approach.

References

- eMunch. Edvard Munchs tekster. Digitalt arkiv* (eMunch. Edvard Munch's Writings. Digital Archive). Accessed March 4, 2017. <http://emunch.no/>.
- 'Ulan. Chéret, Jules'. In *The Getty Research Institute. Union List of Artist Names Online*. Accessed March 4, 2017. http://www.getty.edu/vow/ULANFullDisplay?find=Ch%C3%A9ret&role=&nation=&prev_page=1&subjectid=500030480.
- 'Ulan. Carrière, Eugène'. In *The Getty Research Institute. Union List of Artist Names Online*. Accessed March 4, 2017. <http://www.getty.edu/wNFullDisplay?find=Ch%C3%A9ret&role=&nation=&page=1&subjectid=500007012>.

Crowdfunding the digital scholarly edition

Webcomics, tip jars, and a bowl of potato salad

*Misha Broughton*¹

Paper presented at the DiXiT Convention 'Academia, Cultural Heritage, Society', Cologne, March 14-18, 2016.

Digital Scholarly Editing suffers a funding mismatch. Editors often create editions using grant funding, a model particularly geared to the print paradigm publication model. Namely, these grants provide a set amount of funds, over a limited period, to support the scholarly work which leads to the edition. Under ideal (print) circumstances, by the end of the funding period, a complete edition would be ready to be handed over to a publisher who then handles the expense of typesetting, printing, binding, and distributing the finished edition. In the digital paradigm publishing model, however, publishers are taken out of the equation. And while the publishing is much cheaper – and much faster – under the digital paradigm, and these costs and time can be covered easily during the grant period, digital publications have ongoing post-publication costs which print editions do not: server operation costs, software upgrades, data migration, hosting bandwidth, etc. Further, even this admitted dilemma is based on the very assumption that a given editor gets a grant to create her edition at all. The Digital Humanities have been referred to as ‘an old person’s game,’ in that only established scholars with a track record of successful funding are judged worthy of the risks associated with these new methodologies. Early-career or less prestigious scholars often find themselves wishing for the privilege of fretting about what to do when the grant runs out, as it would mean they had a grant to begin with. In either scenario, paywalls and licensing is, of course, an option, but one of the very impetuses for the move from print to digital scholarly editing was the ability to free the edition from the publication house and its price tags, to demolish obstacles – perhaps especially

¹ mishamikeymonk@gmail.com.

financial obstacles – to access without sacrificing quality. But if we are determined to keep our editions free – both in the ‘free beer’ sense, as well as the ‘free speech’ sense – how do we then finance them? In short, how do we monetize ‘free?’

As luck would have it, we scholarly editors are not the first to run into this particular problem. In the mid- and late-1990s, at the same time when scholarly editors were realizing the tremendous power and affordances of digital publication, cartoonists were realizing the same features for their own work. With relatively affordable web-hosting democratically (if commercially) available, they could share their art and funny pictures with fans far beyond what any but the most exclusive syndication contract could offer. Nor were cartoonists the only content creators to realize this potential: musicians, game designers, animators, illustrators, film-makers, and novelists have all come to realize the power of the internet as a distribution vehicle. However, with neither institutional support nor any form of grant funding, these creators were forced to wrestle with the problem of monetizing ‘free’ long ago. In this presentation, I discuss a few of the methods these creators have applied to tackle the apparent incongruity of generating revenue by giving away their product, with a few examples of each approach, and then remark on their commonalities before addressing the question of whether these approaches are appropriate to the digital scholarly edition. This list of approaches is far from comprehensive – there are nearly as many approaches to monetizing free content as there are creators creating it – and if it leans heavily on examples of webcomic creators, this reflects less their dominance in this enterprise and more my own experience.

Approach 1: *Ye Olde Tip Jar*

As soon as PayPal made it possible for small businesses to accept credit card payments without having tremendous (and expensive) credit card processing contracts, ‘Tip Jar’ links became almost ubiquitous on the landing pages of most webcomics. From customized logo links to PayPal’s default orange lozenge with a blue ‘Donate’ inside it, tip jar links gave many small- to mid-sized webcomics the opportunity to allow their fans to directly contribute to the site’s overhead (and the creator’s personal expenses) even as ad revenues began to dwindle. This approach is now all but extinct, however, because – frankly – it was a TERRIBLE form of crowdfunding. If a creator drew too little attention to their tacit donation appeal, donations became an Everybody/Somebody/Nobody problem: Everybody thought Somebody was donating, so Nobody actually wound up donating. Meanwhile, draw too much attention and readers feel pressured, eroding the very good will that would lead them to donate in the first place. As such, while Tip Jar links still abound, few creators rely on it as a sole – or even primary – form of funding.

Approach 1a: *The Fund Drive*

I mention the fund drive as a subset of the Tip Jar, rather than as its own approach, because it is, in principle, the same approach: ask readers to donate money to support a product they already receive for free. However, the fund drive approach sidesteps the problem of reader burnout on donation requests by limiting them

not in space on the webpage, but rather in time: rather than a standing, subtle request for tips, the fund drive picks a specific, limited time and makes blatant and immediate requests for funds. This approach should be familiar to anyone who listens to American public radio or watches public television, or who uses Wikipedia around December. Notable among webcomics, however, is Randy Milholland's *Something*Postive*. Milholland, who was originally running S*P as a hobby project, even as its readership crested 100,000+ readers daily, received one too many e-mails criticizing the professionalism of a project he did for love and provided for free. In response, Milholland opened a donation link and *dared* his readers to donate an amount equal to his day job's salary, he would immediately turn in his notice to his employer and commit himself full-time to the comic, with – he promised – a concomitant rise in its professional standards, as it would now be a professional endeavor. While Milholland himself considered the gambit merely a snub at the more critical of his readership, the drive hit his goal within a month, and – true to his word – Milholland immediately turned in his notice and became a professional, full-time web cartoonist. For several years thereafter, as the anniversary of the first drive approached, he would remind readers that the previous drive had purchased his time for only a single year and open the drive again. He never failed to reach the goal.

Approach 1b: Pay-What-You-Want

If the fund drive ties the Tip Jar mentality to a specific *temporal period*, the Pay-What-You-Want approach ties it to a *temporal event*: the moment when bits are transferred. While the Pay-What-You-Want model is particularly ill-suited for webcomics, whose most recent content usually is displayed prominently on the landing page and immediately consumed, the PWYW model has found traction in other fields, most notably in the *ubuntu* Linux distribution and on the music site *Bandcamp*. In both of these cases, there is a moment when a bulk file – an operating system ISO or a music MP3 – will be downloaded. And in both of these cases, the host presents the user with a transaction window, allowing them to confirm how much money will change hands with the download. Two features separate the PWYW approach from any other e-commerce transaction: 1) the user can (and, according to Bandcamp, *often will*) choose to pay more than the producer is asking, and 2) the feature that makes this approach a subset of the Tip Jar approach: the user can also choose 'zero.'

Approach 2: Membership and Patronage

Rather than the Tip Jar's sporadic small donations, 'Membership' approaches encourage users to donate a set, non-trivial amount at once, usually to 'purchase access' to a site support club for a period of time. This membership often carries certain perks, such as 'making of' material, community forums, discounts on merchandise, etc. While there have been various individual Membership approaches in the past, the majority of this type of funding have moved to a site called Patreon, which handles billing, processing, and user management for Membership models, allowing the content creator to focus on creating content. While webcartoonists are

far from the only creators on Patreon – a brief browse of selections shows musicians, podcasters, Cosplayers, writers, and film-makers – two particular webcartoonists make particularly good examples. Zach Weinersmith and Jeph Jacques both create webcomics in the high-end of the middle on the popularity (and success) spectrum, *Saturday Morning Breakfast Cereal* and *Questionable Content*, respectively. They are not superstars of the webcomic-verse like Penny Arcade or Sluggy Freelance, but neither are they ‘just someone posting MS-Paint squiggles to a Geo-cities page.’ Both, however, make more than US\$84,000 (each) from their Patreon accounts (which is but one of several revenue streams for their comics). How? In both cases, these donations are from more than 3000 individual backers, averaging less than \$3 per backer.

Approach 3: Merchandise

Among webcomics, merchandise sales tend to fall into two categories: souvenirs emblazoned with the comic’s characters, situations, or dialog, and ‘Dead Tree Editions,’ print versions of the material available online. The former case is, arguably, not terribly appropriate to Digital Scholarly Editions: it is unlikely there will ever be a huge demand for ‘*Letters of 1916 Project*’ t-shirts, or ‘Transcribe Bentham’ ball caps. The latter case, however, bears consideration, not least because at least one DSE, ‘Vincent Van Gogh The Letters,’ already is using almost exactly this approach: making material available for free online while simultaneously selling a print edition.

Nor is this approach limited to webcomics. The novelist Corey Doctorow’s entire publication strategy is to make his novels available, first and foremost, as freely downloadable plain text files on his website *craphound.com*. Doctorow then invites readers who prefer a different format or platform (PDF, ePub, Kindle, etc.) to transform the files as they need, with a request that they then upload the resulting format for use by other readers. He then publishes those same books in a print edition through Tor, North America’s largest science fiction publisher. While Doctorow is reticent to discuss the financial success of this model, it is worth noting that he flatly refuses direct reader monetary donations, preferring instead that consumers purchase a copy for a library in want of one. This way, his publisher is never cut out of the revenue stream. While this is far from proof of his affluence, it is perhaps telling that he can afford to decline direct funds in favour of keeping his publisher invested.

Approach 4: Delivery on Payment

While this approach also encompasses the sites *indiegogo* and *GoFundMe*, *Kickstarter* is by far the most well-known of the sites facilitating Delivery on Payment approaches. Like the subscription lists of 18th and 19th century publishing, Kickstarter collects promises of funding to count against a minimum amount the project needs to complete its smallest production run. Unlike 18th and 19th century subscription lists, however, Kickstarter has both a worldwide outreach (the internet) and a method of painfully billing or returning backers

funds (credit card billing). If the project reaches its funding goal, the project is considered successful, backer's credit cards are charged for their pledges, and funds are dispersed to creators (who are then expected to get busy creating). If the project does not reach its funding goal, the project has failed, and no-one is billed. This business model has led to several unlikely (and noteworthy) success stories, two of which I would like to highlight:

First, in 2014, Zach 'Danger' Brown of Columbus, Ohio submitted a Kickstarter project to create a single bowl of potato salad (for his own consumption), with a minimum funding threshold of \$10. Admittedly only submitting the project to test the bounds of Kickstarter's policy on projects containing 'a product,' Brown was surprised when the project was accepted. However, he was *more* surprised when the campaign received \$500 in pledges in its first day. As the campaign became a viral sensation, pledge totals kept increasing, reaching \$55,000 by the end of the campaign one month later. (It is worth noting that Brown did not wish to keep the funds, but was stymied by another Kickstarter requirement that funds raised not be donated to charity. Revisiting his earlier ingenuity, he noted that the wording said only that funds not be donated DIRECTLY to charity, and that there was no regulation against using extra funds raised to host a potato salad themed festival, the proceeds of which then benefited the Columbus Food Bank.)

Second, in early 2015, three enterprising game developers created a campaign for a card game called 'Exploding Kittens,' seeking US\$10 000 for a minimal print run. Game play was simple, if a bit macabre: 4 players take turns drawing from a deck of 56 specially-created cards. Some of these cards are 'Kittens,' and if a player draws a Kitten card and has no cards with which to distract or otherwise de-escalate the Kitten, then the Kitten inadvertently causes the player's grisly 'death' through its adorable antics, removing that player from the game. The creators, each something of an internet celebrity in his own right, reached their funding goal in an hour. By the end of the first day, they had reached \$1,000,000. At month's end, they had raised over \$8.7M from over 219,000 backers, making 'Exploding Kittens' one of the most successful Kickstarter campaign's in the site's history.

All of these examples, both of Approaches and instances of them, would be useful if we could generalize from those specifics some commonalities by which they succeed. I argue that the most important commonalities of the approaches mentioned here are these three: First, they are brand driven; the creator and/or the title becomes metonymic to the body of work as a whole. Supporters are not paying for the product they have seen already (which they can, and have, accessed without paying), but for the continuation of the brand they enjoy. Second, they are participatory; they offer supporters the opportunity to belong in the production process. Backers of 'Exploding Kittens' do not pay so much for a fairly simple card game, nor do Patreon supporters of Questionable Content have a desperate need for access to the Members-Only forums. Rather, they pay to be able to say 'I was part of this.' Third, they are *extremely* large, and *extremely* lossy; webcomic creators estimate that only 1-3% of readers donate. The fact that these approaches work at all speaks to the *tremendous* number of readers that 1-3% is applied to.

Which brings us to the final, and most important question: Are these approaches appropriate to the funding of Digital Scholarly Editions. To which I answer, quite

simply, no. They are not. Considering the commonalities of these approaches I highlight above, I argue that to build a brand of our creations, one that some segment of the public would be willing to pay in order to participate in it, requires not readers or users or stakeholders, but FANS. And, as a field, we have not been very good at creating fans of our work. In the print paradigm, the publisher often fills this void, creating ‘fans’ out of the libraries, scholars, and instructors who would pay (or make their students pay) to use scholarly editions. But with the publisher cut out of the workflow, the generation of enthusiasm – of fandom – has faltered. And without fans – without a fairly large number of fans, in fact – we have no hope of implementing any of these strategies successfully. If it is true, as one editor I overheard quipped, that the only people interested in scholarly editing are scholarly editors, then we could simply pass a single, crumpled 5 Euro note from hand to hand at the next TEI meeting and call it a day on crowdfunding.

However, I would be remiss if I ended without complicating that ‘no.’ These approaches are not terribly appropriate to scholarly editing *now, as we practice it*. And yet... There is a growing conversation (and practice) of crowdsourcing transcription and tagging in the scholarly editing community, with ‘Transcribe Bentham,’ ‘eMop for Typewrite,’ and ‘The *Letters of 1916* Project’ being notable success stories. Crowdsourcing – like crowdfunding – is an invitation for the public to participate in and to support production, only in this case with their time and labour rather than their funds. Both forms of participation require the same enthusiasm from an interested public. Why should we believe that the same public would be willing to support us with their labor, but not with their funds?

With this in mind, I would like to end on the question ‘How *could* these approaches be appropriate?’ To which I would offer the following, tentative, answers.

First, we must make fans out of our readers and users. Perhaps, in some cases, we have to first make readers and users out of an indifferent public, and then make those readers and users fans. Regardless, the watchword is enthusiasm, and we must be the fountainhead of that enthusiasm.

Second, we must ask. Quite simply, we actually have to try it. While crowdfunding of science research is enjoying a slow ascendancy, very little is being (visibly) done to pursue crowdfunding of digital scholarly editions. Even in the case of merchandising, I said only that it was ‘arguably’ inappropriate for this genre. I cannot say it is absolutely inappropriate simply because, to the best of my knowledge, no-one has tried.

References

Comics and Media

Craphound. Doctorow, Corey. www.craphound.com. Last accessed 7 March 2017.

Questionable Content. Jacques, Jeph. www.questionablecontent.net. Last accessed 7 March 2017.

Saturday Morning Breakfast Cereal. Weinersmith, Zach. www.smbc-comic.com. Last accessed 7 March 2017.

*Something*Positive*. Milholland, Randal K. www.somethingpositive.net. Last accessed 7 March 2017.

Funding and Business

BandCamp. www.bandcamp.com. Last accessed 7 March 2017.

GoFundMe. www.gofundme.com. Last accessed 7 March 2017.

Indiegogo. www.indiegogo.com. Last accessed 7 March 2017.

KickStarter. www.kickstarter.com. Last accessed 7 March 2017.

Patreon. www.patreon.com. Last accessed 7 March 2017.

Software

Ubuntu Linux. www.linux.com. Last accessed 7 March 2017.

Editing medieval charters in the digital age

Jan W. J. Burgers¹

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

For the student of medieval history, charters are a main source of information. This is because charters, being judicial documents issued by one or more authoritative persons, give objective facts about a wealth of subjects, and because they are transmitted in massive quantities – in European archives, hundreds of thousands if not millions of charters are kept, most of them from the late Middle Ages. For these reasons, charters have been edited for as long as medieval history is studied. Those charter editions were often in the form of special collections, the 'charterbooks' (*Urkundenbücher* in German), in which the documents were edited according to a specific diplomatic method.

In the 19th century, German scholars developed this edition method, their most important project being the publication of the charters issued by medieval German kings in the series *Monumenta Germaniae historica*.² In that period, many more charterbooks started to appear, in Germany and elsewhere. All editions consisted of a chronologically arranged series of documents related to a certain person, such as a king or count, or to a certain city, institution or region, as for instance the *Urkundenbuch* concerning the city of Hamburg.³

In the Netherlands, on which I will focus in this paper, the historians were lagging somewhat behind, but in the course of the 20th century they also began

1 jan.burgers@huygens.knaw.nl.

2 In the *Diplomata* series of the *MGH*, the first work of which was edited by G.H. Perz, Hannover, 1872. Since then, 22 titles have appeared, many of them in two or more volumes. The series is published on the internet at <<http://www.dmgh.de/de/fs1/object/display.html>>. (All internet sites mentioned in this article were accessed in October 2017.)

3 Lappenberg – Nürnberg 1842-1967. A digital continuation, up to 1529, is currently made in *Das virtuelle Hamburgische Urkundenbuch*: <http://www.spaetmittelalter.uni-hamburg.de/hamburgisches_ub>.

to make charter editions according to the accepted scholarly method. Here also, *Oorkondenboeken* were published concerning a single region, often a modern province, which in the Middle Ages more or less coincided with a bishopric, duchy or county. The first edition that fully applied the diplomatic method was the charterbook of Holland and Zeeland (Koch et al. – Kruisheer 1970-2005). These editions all start with the earliest documents, but they mostly stop around 1300, when the number of charters explodes. This can be illustrated by the charterbook of Holland and Zeeland: the first volume, spanning the period from the late 7th century to 1222, contains 423 numbers (a number being the edition of a charter text); the fifth and last volume, containing 964 numbers, runs only from March 1291 to November 1299. In other words, the first volume has an average of less than one number per year, while in the fifth volume we have an average of over 120 numbers per year. And after 1300, the number of charters keeps growing. No one has ever counted them, but in the Netherlands alone there must be several hundreds of thousands of charters from the late Middle Ages.

Moreover, editing charters according to the classical scholarly method is a laborious and therefore slow process. Before you can even begin, your source materials must be gathered from all relevant archives and libraries, some of them in far-away places. Moreover, in conjunction with the edition, much diplomatic research has to be done, for instance on the manufacture of the charter and the transmission of its text; the resulting study is printed as an introductory note to each charter that is edited. As a consequence, making a traditional charterbook is a formidable task. The work at the charters of Holland and Zeeland, for instance, started in the 1930s; the first volume appeared only in 1970, the fifth and final volume in 2005.

A solution for speeding up the slow editing process was found in subdividing the corpus, for instance by dividing a province in smaller regions or distinct units such as the charters of a monastery or the deeds of a count. This was done in the Dutch charterbooks of Noord-Brabant (Camps *et al.* 1979-2000) and Guelders (Harenberg *et al.* 1980-2003) and, in Belgium, in the edition of charters issued by the counts of Flanders (Prevenier *et al.* 1966-2009). Another solution to edit the mass of late medieval charters was found in the publication of collections of so-called *regests* (in German: *Regesten*). In those editions, instead of the text of the charter, only a short abstract is printed, together with a minimum of data, such as its archival location.⁴

At the close of the twentieth century, the slow work tempo connected with the edition of charterbooks went out of favour with the decision makers. They still acknowledged the usefulness of those publications, but they wanted faster results. They also realised that by continuing the accepted edition method, the thousands of late medieval charters would in the foreseeable future not become available for the researchers.⁵

⁴ An example in the Netherlands is (Muller 1981).

⁵ In the Netherlands, a first warning came from the experienced diplomatist Prevenier (Prevenier 1985), who objected to the time-consuming and, in his eyes, minimal gains of the *l'art pour l'art* of the classical method of editing charters.

At the same time, the computer started to revolutionise the sciences, including the field of diplomatics. In the course of time, a number of different possibilities of digitally editing charters came forward, which were applied in their pure form or in various combinations.

First of all, one can digitise existing charterbooks. This has been done rather extensively, all over Europe. It is a cheap and quick method, and it makes the contents of those old publications, which often lie hidden in a few specialised libraries, widely available. The simplest method is to scan the pages of a charterbook and put the pictures online. Recently, more advanced methods have been applied, using OCR. The Dutch charterbooks, for instance, have been digitised by the Huygens Institute for the History of the Netherlands in The Hague.⁶ In those editions, one can leaf through the original printed books as an image, as PDF, or even as OCR. Of course it is also possible to do a full text search.

In such a type of digital edition, the original printed book is still the visual reference point. In many of the recently digitalised editions, however, the screen image no longer has a typographical connection with the original publication. For instance the *Württembergisches Urkundenbuch*, running from circa 700 to 1300, is now digitised in a manner in which each charter is displayed on a single screen page.⁷ Many old charterbooks recently have been digitised according to the same principle, such as the editions published by the École des Chartes in Paris.⁸

Secondly, one can simply continue an existing project of a charterbook edition, but now publish the new volumes exclusively in a digital form. As far as I know, this option is not widely used⁹, but an example can be found in another Dutch charterbook, that of Noord-Brabant. Between 1979 and 2000, two volumes (in four books) were published concerning the charters of various regions of this province, but the ongoing work on volumes three and four is now made exclusively available on the internet.¹⁰ From 2010 onward, all completed items are added immediately to the database, which therefore is expanding continually. As a consequence, researchers do not have to wait for a decade or so for a next volume of the charterbook, but the workload has remained the same and the editing process is just as slow as it was before. The new medium has additional advantages: pictures of the source are now added to the edition. Also added is a list of names of persons, places or institutions found in the text, in the form of hyperlinks. The user now has the option to click a link, which gives a list of all other charters in which that name is mentioned. The names are normalised to modern spelling, greatly simplifying the search actions but again adding to the workload of the editor.

6 See for instance the retro-digitised *Oorkondenboek van Holland en Zeeland*: <<http://resources.huygens.knaw.nl/retroboeken/ohz>>.

7 <<https://www.wubonline.de>>.

8 For instance the digitisation of (Guérin – Celier 1881-1958). <<http://corpus.enc.sorbonne.fr/actesroyauxdupoitou>>.

9 An example is the digital continuation of the *Hamburgisches Urkundenbuch* (see above, footnote 2), but this digital continuation is up to now rather fragmentary.

10 <<http://resources.huygens.knaw.nl/oorkondennoordbrabant>>. For the earlier printed edition, see above footnote 5.

A third possibility is to edit a single source, or a single source complex, containing the texts of a great many charters. That way, one is relieved of the time-consuming task of an all-out archival search and a diplomatic study of every single document. Consequently, the editing process is sped up considerably. This method is followed in the digital edition of the *Registers of the counts of Holland* of the period 1299 to 1345.¹¹ There are 22 register volumes extant, together counting some 1000 tightly written pages, containing over 3500 documents, mostly charters issued by the count. In the edition, these texts are presented just as in the classical charterbooks. Attached are scans of all registers, thus offering the possibility to inspect the original source; it is also possible to virtually leaf through every single register volume. Of course, the user can search through the full text of the edition, and here also are at the bottom of each text hyperlinks of the normalised names of persons and places, as well as of the most important terms found in the document. These registers have been edited in circa six years by one single person, working half time, while the work on the traditional *Oorkondenboek van Holland en Zeeland*, containing about the same number of charter texts, took more than sixty full-time man-years.

However, all these modern practices of digital charter editions have not solved the problem of coping with the hundreds of thousands of late medieval charters. A complete edition of all charters still remains out of reach. The only immediate solution to this problem seems to lie in a rough digital presentation of large quantities of materials. This implies a departure from the classical diplomatic method of editing charters. In practice, this means omitting the study of the transmission of each charter text, the identification of textual variants, and often even the text edition itself. In many cases, only a short abstract of the text is included, which makes such type of digital editions the equivalent of the old printed collections of regests.

A first option therefore is to continue those earlier publications of regests with digital means. An example is *Regesta imperii*, the collection of regests of charters issued by the German kings and emperors. This project started in 1829, as a kind of preliminary work for the definitive edition of the royal charters in the *Monumenta Germaniae historica* (Bohmer 1833). New volumes are still being published in book form (now over 90 volumes have appeared, containing some 130,000 numbers), but since 2001 the data are also available online in the form of a searchable database.¹²

A more advanced example is constituted by *Diplomata Belgica*.¹³ This is also on the basis of a regest edition, namely the 11 volumes of the *Table chronologique*, started by Alphonse Wauters in 1866, but it is enriched by materials from various charter editions and by pictures of the original documents. At present, the database contains the metadata of about 35,000 charters up to the year 1250; included are also some 5000 photographs and almost 19,000 full text transcriptions, taken from earlier printed editions. However, it is clear that in this project much work

11 <<http://resources.huygens.knaw.nl/registershollandsegrafelijkheid>>.

12 <<http://www.regesta-imperii.de/en/home.html>>.

13 <<http://www.diplomata-belgica.be>>.

still has to be done by hand, and from the mass of charters after 1250 only a selection will be incorporated.

Another method is to work on the basis of the contents of individual archives. By limiting oneself that way, chronological completeness can be obtained, be it only for a small section of the late medieval documents. A spectacular example is constituted by the *Monasterium* project.¹⁴ Started in Austria in 2002, its database at the moment contains over 838,000 images of 623,000 charters from 164 archival funds in fourteen countries. To make such an effort possible, the project is an ad-hoc collaboration of some 60 archives from central and southern Europe. As a consequence, the metadata are not tailored to one single standard, but simply are adopted as they are supplied by the various institutions. Next to the abstracts, sometimes also transcriptions of the texts or scans of the original charters are added, but especially the transcriptions are often lacking. To emend this defect somewhat, the collaboration of users is sought. They are invited to add those transcriptions as well as other useful elements, under the eye of a watchful moderator. They may even add new items. Thus, *Monasterium* is in fact a digital collection of archival data as well as a steadily growing digital-born charterbook; moreover, also included are digitisations of 137 old printed charterbooks.

Operating along the same lines is the *Cartago* project in the Netherlands.¹⁵ This database contains archival data, many scans and a number of transcriptions from charters in the archives of the northern provinces of Groningen and Drenthe, and recently also from the adjacent German region of Ost-Friesland. Like *Monasterium*, *Cartago* also includes digitised charterbooks, and it too seeks the collaboration of the public. It must be said, however, that up to now this user participation seems not to be very successful, quantitatively nor qualitatively. The database is also not very user friendly, as the search results do not disclose beforehand whether one will get an original charter, or a later transcript, or even a printed text from a modern charterbook.

Finally, at the moment a similar but more encompassing project is being initiated at the Huygens Institute. This aims to be a Dutch national charter portal, the *Digitale Charterbank Nederland* (DCN).¹⁶ In this portal, the data will be collected of all original medieval charters that are kept in Dutch archives. To speed up the working process, it is tried to harvest those data automatically from the online inventories of the various archival funds, by selecting the items containing the Dutch word 'charter' or one of the other archival terms used to designate an original charter. Included in the dataset are the obvious elements: the location and shelf number of the document, the summary of the text as given in the inventory, and, if present, the regest as well as thumbnails of the scans. When *DCN* is completed, the user will be able to search all digitised items in one go; in order to consult the inventory itself or to see the full picture of the document, he or she will be redirected to the website of the archival institution.

14 <<http://monasterium.net/mom/home>>.

15 <<http://www.cartago.nl/nl>>.

16 The *DCN* project is at the Huygens Institute supervised by J. Burgers and dr. Rik Hoekstra.

The advantages of large charter collections such as *Monasterium*, *Cartago* and *DCN* are obvious. Every historian or other interested person will be able to search through the data of thousands of medieval charters, and in many instances even have a picture of the document. However, when compared with the classical charterbooks, the disadvantages are equally clear. Instead of a scholarly edition, including a study on the origin and transmission of the text, the user now will have only the most basic data, often not even a transcription of the charter.

However, even within these minimal data much research can be done from one's armchair that before was impossible to do without travelling to a great number of archives and consulting an even larger number of inventories. Thus, much basic research, for instance on specific persons or places, will be facilitated greatly. But these large digital editions will also allow for more wide-ranging historical studies, for instance on the *longue durée* of many medieval developments, such as feudal institutions, trade and industry, the reclamation and use of farmlands, or the process of substituting Latin for the vernacular in official documents. And when a sufficient number of pictures are available, even the traditional diplomatic and paleographical research can be done.

Finally, it is possible that those pictures will come to additional good use in the foreseeable future. If recent efforts of developing a tool for automatic recognition of handwritten texts proves successful, thousands of – more or less rough – transcriptions could be attached to the data of those large-scale editions, thus in one stroke greatly enlarging the research possibilities. It is even conceivable that the online pictures of medieval charters can play a role in the development of such a reading tool. The scans are already available in large quantities and they are often of sufficient quality. Moreover, charters are uniformly structured documents, consisting of a single block of text. And the script of charters is mostly quite regular and stylised. All this seems to make them ideal training material for a learning machine.

To conclude, the digital age probably has arrived just in time. Just as the traditional method of editing charters had come to a dead end, because of the impossibility of editing the mass of late medieval documents, the computer offered new possibilities. Already various digital methods have been developed for making charters available for the researcher. It is likely that not all ideas will turn out to be equally valid or fruitful, but the best path forward probably lies in integrating the various methods, as is done already in for instance the *Monasterium* portal, where rough archival data are combined with digitised charterbooks and digital-born editions. In the Netherlands, something similar could be done by integrating *Cartago* and *DCN* with the already digitised *Oorkondenboeken* and regests collections, and with digital-born editions such as the *Oorkondenboek Noord-Brabant* and the *Registers of the counts of Holland*.

Whichever path we take, the consequences for the researcher will be profound. The digital edition methods often constitute a break with the classical diplomatic approach, as they omit much information and often even a transcription of the text. Scholars will have to learn to cope with that, and will have to acquire new skills, such as reading old script. But the disadvantages are offset by the massively increased volume of materials and the greatly enhanced search possibilities. If

digital charter editions are taken up and executed with the necessary vigouresness and discrimination, they will give a strong boost to all types of research on the history of the Middle Ages.

References

- Böhmer, J. F. 1833 (2nd ed. Innsbruck, 1889). *Regesta chronologico-diplomatica Karolorum. Die Urkunden sämtlicher Karolinger in kurzen Auszügen*. Frankfurt am Main.
- Camps H. P. H., M. Dillo, G. A. M. Van Synghel (eds.). 1979-2000. *Oorkondenboek van Noord-Brabant tot 1312*. 2 vols. 's-Gravenhage.
- Guérin, P., L. Celier (eds.). 1881-1958. *Recueil des documents concernant le Poitou contenus dans les registres de la Chancellerie de France*. 14 vols. Poitiers.
- Harenberg, E. J., M. S. Polak, E. C. Dijkhof (eds.). 1980-2003. *Oorkondenboek van Gelre en Zutphen tot 1326*. 8 vols. 's-Gravenhage.
- Koch, A. C. F., J. G. Kruisheer, E. C. Dijkhof. 1970-2005. *Oorkondenboek van Holland en Zeeland tot 1299*. 5 vols. 's-Gravenhage etc.
- Lappenberg, J. M., H. Nirnheim (eds.). 1842-1967. *Hamburgisches Urkundenbuch 786-1350*. 4 vols. Hamburg.
- Muller, P. L. 1881. *Regesta Hannonensia. Lijst van oorkonden betreffende Holland en Zeeland uit het tijdvak der regeering van het Henegouwsche huis, 1299-1345, die in het charterboek van Van Mieris ontbreken*. 's-Gravenhage.
- Prevenier, W., Th. de Hemptinne, A. Verhulst (eds.). 1966-2009. *De oorkonden der graven van Vlaanderen 1128-1206*. 4 vols. Brussel.
- Prevenier, W. 1985. 'Bronnen uit de middeleeuwen'. In *Bron en publikatie. Voordrachten en opstellen over de ontsluiting van geschiedkundige bronnen, uitgegeven bij het 75-jarig bestaan van het Bureau der Rijkscommissie voor Vaderlandse Geschiedenis*, ed. by K. Kooimans et al. 's-Gravenhage, 13-27.
- Wauters, Alphonse. 1866-1971. *Table chronologique des chartes et diplômes imprimés concernant l'histoire de la Belgique*. 11 vols. Bruxelles.
- Württembergisches Urkundenbuch. 1849-1913. 11 vols. Stuttgart.

Editing copyrighted materials

On sharing what you can^{1, 2}

*Wout Dillen*³

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

At the moment, the issue of copyright in digital scholarly editing is a big mess. In the first place, this is the case because copyright law is territorial. In other words: something that is completely legal in one country could be illegal in another. On top of that, these laws change over time, and usually only in the wrong direction (Epstein 2002; Goss 2007). And then there are cases where an especially interesting work threatens to enter the public domain, and the copyright holders start to devise mechanisms for keeping their control a little longer – see, for instance, the recent copyright debate regarding the diaries of Anna Frank (Flood 2016). As scholarly editors, we are expected to take all these issues into account while developing a digital edition, especially if we want our work to be as widely accessible as possible. But this is actually very difficult to do because the legal framework we are supposed to follow is so heterogeneous and prone to change. Rather than going into all the ins and outs of copyright law, however, this conference paper is written from a more pragmatic point of view. Say you find yourself in a situation where the materials you want to edit are still protected by copyright. What are you going to do? How

-
- 1 This conference presentation was based on a longer and more detailed treatment of the topic, published in *Digital Scholarship in the Humanities*. This article 'Digital Scholarly Editing within the Boundaries of Copyright Restrictions' by Wout Dillen and Vincent Neyt was first published online on 4 March 2016 by DSH. The article can be consulted at: <https://doi.org/10.1093/llc/fqw011>.
 - 2 The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme FP7/2007-2013/ under ERC grant agreement n°31360911. In addition, the author received funding to participate in this conference from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° 317436.
 - 3 wout.dillen@uantwerpen.be.

can we try to work creatively within the boundaries of those copyright restrictions, to make as much of our academic output available to the largest possible audience?

Background

By exploring how much of their data scholarly editors can still share *within* the boundaries of copyright, this paper responds to the popular sentiment in academia that copyrighting academic output is bad practice, and that the results of scientific research should be made freely available to the general public by default. Good examples of this idea are the series of blog posts and *The Guardian* columns by paleontologist Mike Taylor who claims that ‘Hiding your research behind a paywall is immoral’ (2013), or the decision by the UK’s Higher Education Funding Bodies to introduce ‘an open access requirement in the next Research Excellence Framework’ (Higher Education Funding Council for England, 2015). Although these examples were written mainly with more traditional academic output formats in mind, they can be applied easily to the academic output of scholarly editors, the most recent realization of which is the digital scholarly edition.

And indeed, at the ADHO conference in Lincoln Nebraska in 2013, scholarly editor Peter Robinson posited a list of desiderata for scholarly editions, that included the dictum that ‘All materials (in a scholarly edition) should, by default be available by a Creative Commons share-alike license’ (2013). Commendable though Robinson’s ambitious program may be, scholarly editors who work with modern manuscripts for which today’s copyright and intellectual property laws are still considerable issues may feel the need to nuance this statement. While I certainly agree that Robinson’s desideratum is a laudable goal that should be pursued whenever possible, the fact remains that it is often unattainable. By arguing that all digital scholarly editions should by default be made available under a CC-BY SA license, Robinson neglects the fact that this license is not necessarily the editor’s to give. While this may prevent the data from being re-used outside the walls of the edition, it should not diminish the value of the data, nor of the functionality that was built around those data. To illustrate this point, the paper zooms in on the practices of a couple of projects that work with copyrighted materials, to see how they deal with this situation.

Curation

The first step in the creation of any scholarly edition – be it digital or in print – is to find an interesting corpus of texts to analyse. Obviously, the ideal situation here would be that the texts of the source materials have entered the public domain, and that high-resolution scans of these materials have been made freely available under a Creative Commons license (see for instance Pierazzo and André 2012). But for scholarly editors working with even more recent materials, conditions are usually less favourable. When copyright restrictions come into play, careful negotiations with authors or the executors of their estates become a crucial aspect of the contractual agreements between what is already a large group of people with

a wide range of (commercial and non-commercial) interests: scholars, memory institutions, publishers, and funders.

The same holds true for the *Beckett Digital Manuscript Project* (BDMP; *Samuel Beckett. Krapp's Last Tape / La Dernière Bande: a digital genetic edition*) for example. At the moment, the BDMP has secured a contract regarding the acquisition of scans with a number of holding libraries, and hopes to strike up even more, similar collaborations in the future. These agreements allow the project to request high resolution scans of the necessary documents for scientific purposes, and to incorporate them into the edition. As with most digital editions, these facsimiles form a crucial part of the BDMP's publication. But what is of course even more important for the edition is the text those facsimiles contain. And the copyright of those texts currently belongs to the Beckett Estate – where it will remain for quite some time to come.

This is what makes a good relationship and workable contracts between Digital Scholarly Editing projects and authors' estates (or their representatives) so important: if reports of troubled relations between scholarly editors and the Joyce Estate for instance teach us anything, it is that a project on such a scale covering an important contemporary author simply would not be possible without it. Thankfully, in the case of the BDMP, all parties involved have realized that the future of scholarly editing is digital, and that the scholarly augmentation of Beckett's legacy will only increase the interest in his works – academic or otherwise. That is why the Beckett Estate agreed to give the directors of the BDMP the license to publish their genetic edition of Beckett's manuscripts online, as long as this happens behind a paywall that pays for the license. Still, *limited* access is better than *no* access; and because the BDMP's contractual agreement with holding libraries and the Beckett Estate stipulates that each of the collaborating institutions is granted institutional access to the edition, this means that there is still a considerable group of people who can access the edition free of charge.

Public Access without a Public License

A different approach to this problem is the one that the *Woolf Online* project has taken (Woolf 2013). This project aims to offer a genetic edition of Virginia Woolf's *To the Lighthouse* in the form of a digital archive that combines transcribed facsimiles of Woolf's manuscripts, a number of the work's editions, extracts from diaries and letters, photo albums, critical reviews, etc. – all of which can be accessed by anyone who visits the project's website. Obtaining the license to publish these materials is in itself already quite an achievement, since the edition incorporates a range of different kinds of materials belonging to a number of different copyright holders. But we have to keep in mind that scholarly edition does not meet the requirements of Robinson's third desideratum either. Although the Woolf Online project has acquired a license to offer its users free access to its materials, those materials are not allowed to leave the project's website.

As it is clearly stated on the project's Copyright Notice, the 'material is provisioned for online publication and reading only at Woolf Online and may not be copied, distributed, transmitted or otherwise altered or manipulated without

the express permissions' of the copyright owners.⁴ This means that the edition's users will not be able to publish the project's data in a new interface, as in the reusable future Robinson envisions. But the question remains how important this final step is to enable further research. Developing a Digital Scholarly Edition, scholarly editors will want to build the best possible environment for their readers to fully appreciate the nuances of the materials they are editing, and to distribute it in such a way that it may be a useful foundation for others to build their research on. To achieve this goal, Creative Commons licenses are useful, but not strictly necessary.

Sharing What You Can

It can be recommended, however, to make academic data that is not bound to copyright restrictions publicly available under as public a license as possible. The most obvious example in this category is metadata. Short of copying their contents, researchers are allowed (and could be encouraged) to describe the resources they are studying in as much detail as they desire, and to share their findings – preferably in a standardized format like RDF. But we can do even more. For researchers who are working with similar data, the steps we take to achieve our results can be just as valuable as the results themselves. Like metadata, a project's documentation, for example, can contain a wealth of harmless information about the project that may be of use for other researchers. In the case of Digital Scholarly Editions, the TEI for example already provides a standard format to share this information: the ODD (or: One Document Does it all). A file that combines a TEI-XML validation schema with human readable schema documentation in a single XML file.

Recognizing the potential of this kind of information for other scholarly editing projects, the BDMP recently put its own documentation online for consultation in the form of a digital Encoding Manual.⁵ The basis of this documentation was a cheat-sheet designed by my colleague Vincent Neyt that helped the project's contributors transcribe Beckett's manuscripts. In collaboration with Vincent, I have expanded this document to include validation information, more examples and explanatory text, more information on encoding practices in general, and the BDMP's ODD validation schema. This way, the project's collectively cultivated expertise does not go to waste, but may help to support (or even initiate) other research instead.

Fair Use

The above was written from a position that takes an absolute view on copyright: namely that nothing that is not protected through copyright may be copied and distributed in any way without the express permission of the copyright holder. But that is of course not exactly true: there are exceptions to the rule that make it possible to share copyrighted materials (to a certain extent), notably by means of

⁴ <http://www.woolfonline.com/?node=about/copyright>.

⁵ <http://uahost.uantwerpen.be/bdmp/>.

the fair use doctrine. The problem with this exception, however, is that it is open to interpretation and assessed on a case by case basis, taking a non-exhaustive list of vague criteria into account.⁶ Still, there are some basic rules of conduct that can be followed to help minimize the risk of litigation to an acceptable degree. For example, using only the most relevant passages of a copyrighted work strictly for research purposes in a non-commercial environment can already go a long way towards convincing copyright holders that the odds would be against them if they decided to pursue the matter legally. This is a strategy we have tried to apply for the *Lexicon of Scholarly Editing* we are developing at the Centre for Manuscript Genetics.⁷

Instead of writing its own definitions, this lexicon aggregates citations in academic outputs (journal articles, monographs, etc.) that are relevant to each of its entries. This way, the passages quoted in the Lexicon can be used to develop a better understanding of certain problematic concepts; to discover what those concepts are called in different languages; and to help textual scholars develop more nuanced arguments in their own writing. In order to publish these citations under the purview of the fair use doctrine, we try to make sure that: (1) the materials are used for research purposes; (2) we only use small fragments – usually only a fraction of the full text; (3) these fragments are attributed consistently to their rightful owner; (4) the use of the fragments only furthers their original aims: namely to further scholarly research; and (5) all of the above is published in a transparent, non-commercial research environment.

Conclusion

In conclusion, I would argue that scholarly editors often still have plenty of room within the boundaries of copyright restrictions for the publication their Digital Scholarly Edition, and all of the ancillary data they produce along the way. This means that the limitations posed on copyrighted materials should not be a reason for academics to stop using those materials. Quite to the contrary: they should be an incentive to make them available for further research by any means possible. There is no reason to put off trying to answer the research questions we face today – especially if we already have the technology to do so. If that means forgoing a CC BY-SA license, or setting up a paywall, so be it. In that case, editors will have to accept the challenge to convince their readers that their corpus is worth consulting in spite of those limitations.

6 <http://www.copyright.gov/title17/92chap1.html#107>.

7 <http://uahost.uantwerpen.be/lse/>.

References

- Epstein, Richard. 2002. 'The Dubious Constitutionality of the Copyright Term Extension Act.' *Loyola of Los Angeles Law Review* 36, 123-158. Accessed April 8, 2016. http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=2359&context=journal_articles.
- Flood, Alison. 2016. 'Anne Frank's diary caught in fierce European copyright battle'. *The Guardian, January 18*. Accessed April 8, 2016. <http://www.theguardian.com/books/2016/jan/18/anne-franks-diary-caught-in-fierce-european-copyright-battle>.
- Goss, Adrienne K. 2007. 'Codifying a Commons: Copyright, Copyleft, and the Creative Commons Project.' *Chi. -Kent Law Review* 82: 963-996. Accessed April 8, 2016. <http://scholarship.kentlaw.iit.edu/cgi/viewcontent.cgi?article=3609&context=cklawreview>.
- Pierazzo, Elena, and Julie André. 2012. *Autour d'une séquence et des notes du Cahier 46: Enjeu du codage dans les brouillons de Proust*. Accessed April 8, 2016. http://research.cch.kcl.ac.uk/proust_prototype/index.html.
- Robinson, Peter. 2013. '5 Desiderata for Digital Editions/Digital Humanists Should Get Out Of Textual Scholarship'. *Paper presented at the DH2014, Lincoln, Nebraska*, July 19. Accessed April 8, 2016. <http://www.slideshare.net/PeterRobinson10/peter-robinson-24420126>.
- Samuel Beckett. *Krapp's Last Tape / La Dernière Bande: a digital genetic edition*. 2015, edited by Dirk Van Hulle and Vincent Neyt (The Beckett Digital Manuscript Project, module 3). Brussels: University Press Antwerp (ASP/UPA). Accessed April 20, 2016. <http://www.beckettarchive.org>.
- Taylor, Mike. 2013. 'Hiding your research behind a paywall is immoral'. *The Guardian, January 17*. Accessed April 8, 2016. <https://www.theguardian.com/science/blog/2013/jan/17/open-access-publishing-science-paywall-immoral>.
- Woolf Online*. 2013, edited by Julia Briggs and Peter Schillingsburg. Accessed April 20, 2016. <http://www.woolfonline.com/>.

What you c(apture) is what you get. Authenticity and quality control in digitization practices¹

Wout Dillen²

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Regardless of whether documents are taking a more prominent place in the edition (see Gabler 2007; Robinson 2013; Pierazzo 2014), it seems inevitable that today's scholarly editors are working more and more with these documents' digitized facsimiles. This may happen already in the research stage of the editing process when digital surrogates are integrated into the editor's workflow and effectively become the documents on which the editor bases her interpretation of the text and of its transmission over time. While these facsimiles are not (and probably indeed should not) be used exclusively, *in lieu* of the original analog source materials, it is nevertheless the case that they come more and more to the forefront as the basis of the editor's analysis and transcription – *precisely because* they are more transportable and durable, and because today's imaging hardware and software is capable of revealing aspects of the documents that are difficult or even impossible to detect with the naked eye.

And as important as these facsimiles have become for the editor, they are even more important for the user. Once they take their place in the digital scholarly edition, facsimiles are the closest the user will get to an unedited representation of the document's physical and textual features – even though, as Hans Zeller

1 The research leading to these results was conducted as part of the author's work on the 'Digital Scholarly Editing and Memory Institutions' project at the University of Borås (Sweden). This is an Experienced Researcher position that is part of the DiXiT network, a Marie Curie ITN which has received funding from the People Programme (Marie Skłodowska-Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n° 317436. This funding also allowed the author to participate in the conference.

2 wout.dillen@uantwerpen.be.

already suggested in the 1970s in his influential essay on ‘Befund und Deutung’, a facsimile is not completely unedited or objective (1971). As the only available visual representative of the physical document, the facsimile becomes the document that the user will use to assess the editor’s transcriptions, claims, and arguments about the text. This means that both users and editors are placing a lot of trust in the digital avatars of physical, historical documents they claim to study. Therefore, this paper argued that issues of authenticity and quality control for digital facsimiles are important issues that need to be addressed at the outset of any digital scholarly editing project. Because too often, we take these digital surrogates at face value.

As editors, we pride ourselves on knowing the difference between the two. We know that text is transmitted from document to document, from medium to medium, and that we should consider the effect that the medium has on the text that is transmitted. Documentary aspects like ‘size’, ‘texture’, ‘coloring’, etc. can help us determine how the text was read at the time the document was made. As textual scholars we are very aware that the medium that carries the text has a very specific impact on our interpretation of that text. And we know that by digitizing an analog object, we effectively are transporting it onto a new medium, which inevitably will have an important impact on our understanding of that text. Some aspects, such as the text, typically will be retained – although the readability of that text will depend strongly on the quality of the images. Some aspects, such as the correctness of its colours, can only be hinted at or assessed in relative terms. And some aspects, like the document’s texture, feel, smell, etc. will be lost. In other words, by transporting select aspects of the original object into a new medium, we are creating a new document, that acts as an intermediary between the original on the one hand, and our edited version on the other.

But even as scholarly editors, do we really treat this new, intermediary document with the same due diligence and scrutiny as we do the documents it mediates between? Personal experience tells me otherwise. Working on the Beckett or Brulez editions at the University of Antwerp’s Centre for Manuscript Genetics as part of my PhD, I was transcribing and checking transcriptions of image files that I had never seen the original of. Of course, I was not the real ‘editor’ of those specific modules. And if I were, I would consider it my editorial duty to go to the archives and see the original documents for myself. But even if that were the case, I doubt that I would spend the same time analysing the ins and outs of every page with the same attention to detail as I did when I was transcribing the documents. In many cases such scrutinous inspection is not even possible, because we have to be very careful not to damage the documents while we are studying them. That is one of the reasons why we digitize in the first place. And when I am using someone else’s scholarly edition, I will think about these issues even less, and interact with the facsimiles ‘as if they were the actual documents’ myself.

This is the premise of my current project at the University of Borås, titled ‘Digital Scholarly Editions and Memory Institutions’ that aims to address these issues by calling for a critical assessment of digitization practices. To do so, the project focuses on the moment the analog object is captured – a process that retains as well as discards a substantial amount of information. Therein lies the meaning of this paper’s title. For the facsimiles that we use in our editions, what you see usually

is what you get: you see an image that you try to make sense of in relation to the edited text, an image that promises to be a faithful representation of the source document, but that is rendered without any data or context to back this claim up. And what you see, is really what you capture: those aspects of the document that are possible to represent in the digital medium as we know it – and really only a selection of those aspects, relative to the means and tools the photographer has at her disposal, the standards that she is required to follow, and the decisions that she has to make. It is these issues and the variables they introduce into the digitization process that this project wants to address.

As the project's first case study, I visited the National Library of Sweden in Stockholm to investigate how this institution handles digitization practices: which standards are used, and why?; how are these standards established?; how are they negotiated between different parties?; and how minutely are they followed in practice? The National Library of Sweden was selected for this case study because 1) it has for some time now been actively involved in the development of a detailed digitization strategy, and 2) it prides itself on putting trained photographers in charge of capturing (i.e. digitizing) the source materials. These factors, as well as some preliminary meetings with Lars Björk – the head of conservation at the National Library – suggested that this institution has a relatively high awareness of the issues this project means to investigate, and that its expertise in these areas may be transferred to memory institutions that have a less advanced digitization strategy. The case study would exist of three parts: 1) a document study – including documents with technical information and digitization manuals that the library had drafted as part of its digitization strategy; 2) an observation study – where I followed a photographer for a couple of days while she was capturing the materials, to study how her workflow was executed in practice, as well as her interaction within this workflow with her hardware and software, and with her colleagues; and 3) a series of interviews with people from the institution who were involved in the digitization process in some form or other.

This paper reported on the preliminary results of this first case study at the KB in Stockholm. By mapping the interactions between different agents who are involved in the digitization process, it aimed to refute the notion that digitization is a simple and straightforward process. Instead, the research suggests that different agents put different demands on the digital object – demands that need to be taken into account by the photographer during the moment of capture. And that even when these issues and agreements are taken into account, digitization often is still a process of 'problem solving'. To this end, the paper presented a set of examples to support the claim that in many cases the quality of the digitization will depend highly on the photographer's professional skills and interpretation (arguably much in the same way as the construction of the 'text' of a scholarly edition depends on the editor's professional skills and interpretation). This would suggest that (as in scholarly editing) the 'accountability' of the photographer is much more important and relevant than her 'objectivity' – and that in a scholarly edition, we should try to find a way to incorporate that accountability into the edition, and thereby raise the user's awareness of the degree of interpretation that goes into the process of creating the digital facsimile.

To conclude, the paper introduced the next proposed case studies that will serve to put the findings at the National Library in Sweden in perspective. By bringing these case studies together, I want to examine how aware these memory institutions are of each other's efforts in this area, how their standards are communicated, and whether they are negotiated further on the international level as well. I think that this mapping of the way in which quality measures are negotiated between different agents is an important first step towards getting a better understanding of the relation between the source document and the digital facsimiles that we are displaying in our editions. And I think that this awareness is essential if we really want to be accountable for all the aspects of our edition, and our interpretation of its materials.

References

- Gabler, Hans Walter. 2007. 'The Primacy of the Document in Editing.' *Ecdotica* 4: 197-207. Accessed 24 October 2016. https://www.academia.edu/166634/The_Primacy_of_the_Document_in_Editing.
- Pierazzo, Elena. 2014. 'Digital Documentary Editions and the Others.' *Scholarly Editing: The Annual of the Association for Documentary Editing* 35: 1-23.
- Robinson, Peter. 2013. 'Towards a Theory of Digital Editions.' *Variants* 10: 105-31. Accessed on 24 October 2016. https://www.academia.edu/3233227/Towards_a_Theory_of_Digital_Editions.
- Van Hulle, Dirk, and Peter Shillingsbiurg. 2015. 'Orientations to Text, Revisited.' *Studies in References* 59. 1: 27-44.
- Zeller, Hans. 1971. 'Befund Und Deutung. Interpretation Und Dokumentation Als Ziel Und Methode Der Edition.' In *Texte Und Varianten. Probleme Ihrer Edition Und Interpretation*, edited by Hans Zeller and Gunter Martens, 45-89. München: CH Beck.

The journal al-Muqtabas between Shamela.ws, HathiTrust, and GitHub

Producing open, collaborative, and fully-referencable digital editions of early Arabic periodicals – with almost no funds

Till Grallert¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016)

The problems at hand

In the context of the current onslaught cultural artefacts in the Middle East face from the iconoclasts of the Islamic State, from the institutional neglect of states and elites, and from poverty and war, digital preservation efforts promise some relief as well as potential counter narratives. They also might be the only resolve for future education and rebuilding efforts once the wars in Syria, Iraq or Yemen come to an end. But while the digitisation of archaeological artefacts recently has received some attention from well-funded international and national organisations, particularly vulnerable collections of texts in libraries, archives, and private homes are destroyed without the world having known about their existence in the first place.²

¹ grallert@orient-institut.org.

² For a good example of crowd-sourced conservation efforts targeted at the Armenian communities of the Ottoman Empire see the Houshamadyan project (<http://www.houshamadyan.org/>), which was established by Elke Hartmann and Vahé Tachjian in Berlin in 2010 and launched an 'Open Digital Archive' in 2015. Other digitisation projects worth mentioning are the Yemen Manuscript Digitisation Project (<http://ymdi.uoregon.edu/>, University of Oregon, Princeton University, Freie Universität Berlin) and the recent 'Million Image Database Project' of the Digital Archaeology Institute (<http://digitalarchaeology.org.uk/>, UNESCO, University of Oxford, government of the UAE) that aims at delivering 5000 3D cameras to the MENA region in spring 2016.

Early Arabic periodicals, such as Butrus al-Bustānī's *al-Jinān* (Beirut, 1876-86), Ya'qūb Ṣarrūf, Fāris Nimr, and Shāhīn Makāriyūs' *al-Muqtaṭaf* (Beirut and Cairo, 1876-1952), Muḥammad Kurd 'Alī's *al-Muqtabas* (Cairo and Damascus, 1906-18/19) or Rashīd Riḍā's *al-Manār* (Cairo, 1898-1941) are at the core of the Arabic renaissance (*al-nahḍa*), Arab nationalism, and the Islamic reform movement. These better known and – at the time – widely popular journals do not face the ultimate danger of their last copy being destroyed. Yet, copies are distributed throughout libraries and institutions worldwide. This makes it almost impossible to trace discourses across journals and with the demolition and closure of libraries in the Middle East, they are increasingly accessible to the affluent Western researcher only.³

Digitisation seemingly offers an 'easy' remedy to the problem of access and some large-scale scanning projects, such as Hathitrust⁴, the British Library's 'Endangered Archives Programme' (EAP), MenaDoc or Institut du Monde Arabe produced digital facsimiles of numerous Arabic periodicals. But they come with a number of problems, namely interfaces not adapted to Arabic script (Arabic books are frequently presented back to front), the absence of reliable bibliographic metadata, particularly on the issue level, and a searchable text layer. Due to the state of Arabic OCR and the particular difficulties of low-quality fonts, inks, and paper employed at the turn of the 20th century, these texts can only be digitised reliably by human transcription (*cf.* Märgner and El Abed 2012).⁵ Funds for transcribing the tens to hundreds of thousands of pages of an average mundane periodical are simply not available, despite of their cultural significance and unlike for valuable manuscripts and high-brow literature. Consequently, we still have not a single digital scholarly edition of any of these journals.

On the other hand, grey online-libraries of Arabic literature, namely *al-Maktaba al-Shāmila*, *Mishkāt*, *Ṣaydal-Fawā'id* or *al-Waraq*⁶, provide access to a vast body

3 In many instances libraries hold incomplete collections or only single copies. This, for instance, has caused even scholars working on individual journals to miss the fact that the very journal they were concerned with appeared in at least two different editions (e.g. Glaß 2004; see Grallert 2013 and 2014).

4 It must be noted that the US-based HathiTrust does not provide public or open access to its collections even to material deemed in the public domain under extremely strict US copyright laws when users try to connect to the collection from outside the USA. Citing the absence of editors able to read many of the languages written in non-Latin scripts, HathiTrust tends to be extra cautious with the material of interest to us and restricts access by default to US-IPs. These restrictions can be lifted on a case-by-case basis, which requires at least an English email conversation and prevents access to the collection for many of the communities who produced these cultural artefacts; see https://www.hathitrust.org/access_use for the access policies.

5 For the abominable state of Arabic OCR even for well-funded corporations and projects, try searching inside Arabic works on Google Books or HathiTrust. The 'Early Arabic Printed Books' (EAPB) project (<http://gale.cengage.co.uk/arabic>), currently under development by GALE in collaboration with the British Library, makes repeated claims of employing 'newly developed optical character recognition software (OCR) for early Arabic printed script' (see this factsheet: http://gale.cengage.co.uk/images/EAPB-Factsheet_English_WEB.pdf). But since they share neither the text layer nor the software their claims cannot be verified. As a substantial number of the digitised books in EAPB are written in languages other than Arabic that employ Arabic script (such as Farsi, Urdu or Ottoman Turkish) and as some works resemble complex manuscripts with multiple commentaries around a main text fully automated text-retrieval is highly unlikely.

6 See <http://www.shamela.ws/>, <http://almeshkat.net/>, <http://saaid.net/> and <http://www.alwaraq.net/>.

of mostly classical Arabic texts including transcriptions of unknown provenance, editorial principals, and quality for some of the mentioned periodicals. In addition, these grey 'editions' lack information linking the digital representation to material originals, namely bibliographic meta-data and page breaks, which makes them almost impossible to employ for scholarly research.

Our proposed solution

With the open scholarly digital edition of *Majallat al-Muqtabas*⁷ we want to show that one can produce scholarly editions that offer solutions for most of the above-mentioned problems – including the absence of expensive infrastructure – through re-purposing well-established open software platforms and by combining the virtues of immensely popular, but non-academic (and, at least under US copyright laws, occasionally illegal) online libraries of volunteers on the one hand with academic scanning efforts as well as editorial expertise on the other. To this end, we transform digital texts from *shamela.ws* into TEI XML, add light structural mark-up for articles, sections, authors, and bibliographic metadata, and link each page break in the digital text to digital facsimiles provided through EAP and HathiTrust; the latter step, in the process of which we also make first corrections to the transcription, though trivial, is the most labour-intensive, given that page breaks commonly are ignored by *shamela.ws*'s anonymous transcribers. The digital edition (TEI, markdown, and a web-display) is then hosted as a GitHub repository with a CC BY-SA 4.0 licence for reading, contribution, and re-use.⁸

We argue that by linking facsimiles to the digital text, every reader can validate the quality of the transcription against the original. We thus remove the greatest limitation of crowd-sourced or grey transcriptions and the main source of disciplinary contempt among historians and scholars of the Middle East. Improvements of the transcription and mark-up can be crowd-sourced with clear attribution of authorship and version control using .git and GitHub's core functionality. Such an approach as proposed by Christian Wittern (2013) has recently seen a number of concurrent practical implementations such as project GITenberg⁹ led by Seth Woodworth, Jonathan Reeve's Git-lit¹⁰, and others.

To ease access for human readers (the main projected audience of our edition) and the correction process, we provide a basic web-display that adheres to the principles of GO::DH's Minimal Computing Working group.¹¹ This web-display is implemented through an adaptation of the TEI Boilerplate XSLT stylesheets to the needs of Arabic texts (or right-to-left writing systems in general) and the parallel display of facsimiles and the transcription. Based solely on XSLT 1 and

7 For a history of Muḥammad Kurd 'Alī's journal *al-Muqtabas* (The Digest) see Seikaly (1981) and the readme.md of our project's GitHub repository: <https://github.com/tillgrallert/digital-muqtabas>.

8 The text of *al-Muqtabas* itself is in the public domain even under the most restrictive definitions (i.e. in the USA); the anonymous original transcribers at *shamela.ws* do not claim copyright; and we only link to publicly accessible facsimile's without copying or downloading them.

9 <https://gitenberg.github.io/>.

10 <https://github.com/Git-Lit/git-lit>.

11 <https://go-dh.github.io/mincomp>.

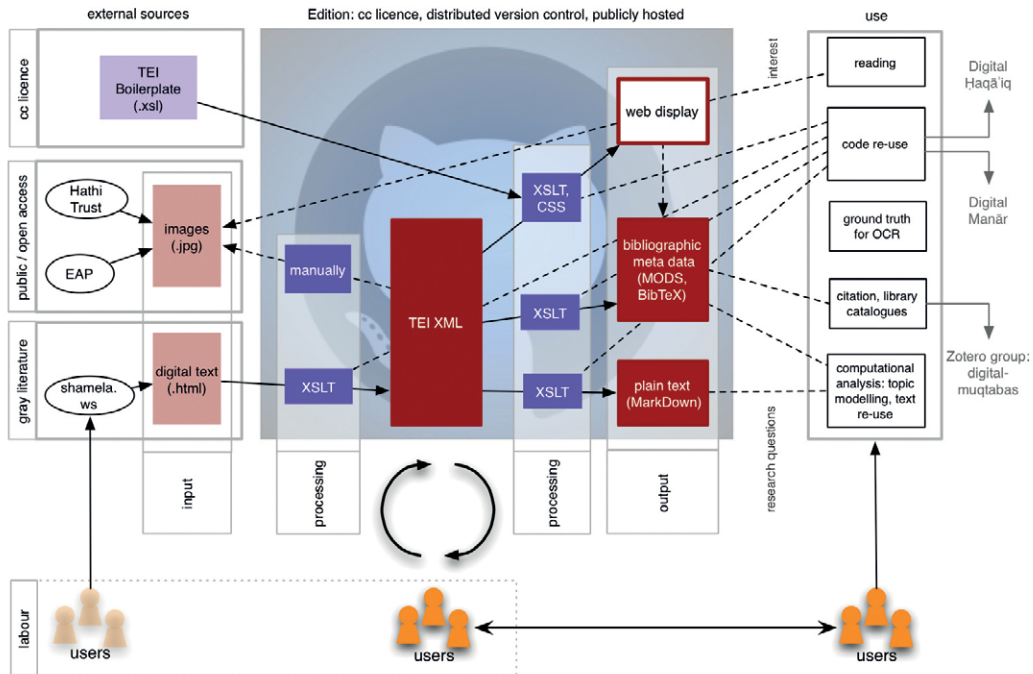


Figure 1: The journal *al-Muqtabas*: Project scheme.

CSS, it runs in most internet browsers and can be downloaded, distributed and run locally without any internet connection – an absolute necessity for societies outside the global North.

In addition to the TEI XML files we provide automatically generated structured bibliographic metadata for every article in *al-Muqtabas* as MODS (Metadata Object Description Schema) and BibTeX files that can be integrated easily into larger bibliographic information systems or individual scholars' reference managing software. Currently, we provide stable URLs to all elements of the mark-up by combining the base-url of the TEI files with their @xml:ids. For future iterations, we plan to make the XML referenceable down to the word level for scholarly citations, annotation layers, as well as web-applications through a documented and persistent URI scheme such as Canonical Text Services (CTS) URN (*cf.* Kalvesmaki 2014). To further improve access to individual articles and allow for a search of bibliographic metadata across issues and beyond GitHub's search functionality we feed our MODS files into a public Zotero group.

In order to contribute to the improvement of Arabic OCR algorithms, we will provide corrected transcriptions of the facsimile pages as ground truth to interested research projects starting with transkribus.eu.

Finally, by sharing all our code, we hope to facilitate similar projects and digital editions of further periodicals. For this purpose, we successfully tested adapting



Figure 2: Web display of *al-Muqtabas* 6(2).

the code to ‘Abd al-Qādir al-Iskandarānī’s monthly journal *al-Ḥaqā’iq* (1910-12, Damascus)¹² in February 2016.

Conclusion

Cultural artefacts, and particularly texts, face massive challenges in the Middle East. We propose a solution to some of these problems based on the principles of openness, simplicity, and adherence to scholarly and technical standards. Applying these principles, our edition of *Majallat al-Muqtabas* improves already existing digital artefacts and makes them accessible for reading and re-use to the scholarly community as well as the general public. The paper discusses the particular challenges and experiences of this still very young project (since October 2015).

¹² <https://github.com/OpenAraPE/digital-haqaiq>; on the history of *al-Ḥaqā’iq* and some of its quarrels with *al-Muqtabas* see Commins (1990, 118-122).

References

- Commins, David. 1990. *Islamic Reform: Politics and Social Change in Late Ottoman Syria*. Oxford: Oxford University Press.
- Glaß, Dagmar. 2004. *Der Muqtaṭaf und seine Öffentlichkeit. Aufklärung, Raisonement und Meinungsstreit in der frühen arabischen Zeitschriftenkommunikation*. Würzburg: Ergon Verlag.
- Grallert, Till. 2013. 'The Puzzle Continues: Al-Muqtaṭaf Was Printed in Two Different and Unmarked Editions.' *Blog post. Sitzextase. August 19*. <http://tillgrallert.github.io/blog/2013/08/19/the-puzzle-continues/>.
- Grallert, Till. 2014. 'The Puzzle Continues II: In Addition to Al-Kabīr and Al-Ṣaghīr, Al-Muqtaṭaf Published Slightly Different Editions in Beirut and Kairo.' *Blog post. Sitzextase. January 19*. <http://tillgrallert.github.io/blog/2014/01/19/the-puzzle-continues-2/>.
- Kalvesmaki, Joel. 2014. 'Canonical References in Electronic Texts: Rationale and Best Practices.' *Digital Humanities Quarterly* 8 (2).
- Märgner, Volker, and Haikal El Abed (eds). 2012. *Guide to OCR for Arabic Scripts*. London: Springer. <http://link.springer.com/book/10.1007/978-1-4471-4072-6>.
- Seikaly, Samir. 1981. 'Damascene Intellectual Life in the Opening Years of the 20th Century: Muhammad Kurd 'Ali and Al-Muqtabas.' In *Intellectual Life in the Arab East, 1890-1939*, edited by Marwan Rafat Buheiry, 125-53. Beirut: American University of Beirut.
- Wittern, Christian. 2013. 'Beyond TEI: Returning the Text to the Reader.' *Journal of the Text Encoding Initiative 4: Selected Papers from the 2011 TEI Conference*. <http://jtei.revues.org/691>.

Digital editions of artists' writings

First Van Gogh, then Mondrian

Leo Jansen¹

Opening keynote given at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

In October, 2009, the results of the Van Gogh Letters Project were published online: www.vangoghletters.org. In this talk I will look back on the history of the project. When and why was it started, under what circumstances and with what aims; why did it take so long, what changed over the course of 15 years between start and finish, which methodical and practical changes occurred?

On the face of it there is nothing exceptional in publishing the correspondence of a historical person; there are hundreds if not thousands of scholarly editions that make letters available for reading and studying. On the other hand, as I hope to show, in the case of Van Gogh's correspondence there were specific circumstances that turned this into not just another edition, or another scholarly edition for that matter. To close, I will briefly sketch the outlines of a project that could be considered the Van Gogh project's successor, the Mondrian Edition Project.

Why?

Before embarking on an ambitious edition project there is a fundamental question that needs a very convincing answer before one can start to think further about it. This question is: why should we do this? In general, the answer depends on two conditions. 1: the historical importance of the text; and 2: the possible contribution of the projected edition to scholarship in the related academic disciplines. Once the

¹ leo.jansen@huygens.knaw.nl.

importance has been substantiated convincingly and it can be argued that scholars will benefit substantially, one can start to think about what this publication might look like and what is the best way to realize it is.

So why publish Vincent van Gogh's correspondence? Are these letters so important? There are 820 letters from Vincent van Gogh still extant and 83 he received. They date from 1872 till 1890, the year of his death. About two-thirds of his letters were written to his brother and confidant Theo. The rest is written to his sister Will, artist-friends such as the Dutch Anthon van Rappard and Frenchmen like Paul Gauguin and Emile Bernard, and a few others. Roughly two-thirds are written in Dutch, one-third in French. If we limit ourselves to Vincent van Gogh's own letters, they are an invaluable source to those who seek deeper insight in the art, the ideas and the biography of one of the most important artists of modern art. This justifies their publication beyond any reasonable doubt. The fact that according to many they have literary merits as well is an extra argument to publish them.

The next question, then, is: what is the type of edition that best suits the demands of this twofold readership, the scholars on one side and the more literary interested readers on the other? The publication history of the letters during the decades before we started the project shows that there were already different types of publications available in many languages, ranging from complete editions to all sorts of anthologies and selections. It is important to realize, however, that most of these publications were edited by dedicated individuals who were not trained in textual criticism or scholarly editing.

And so the Van Gogh Museum decided to come to terms with the needs of art-historians worldwide. In 1994 the Van Gogh Letters Project was launched as a collaborative venture under the auspices of the Van Gogh Museum in Amsterdam and the Huygens Institute for the History of the Netherlands in The Hague. This was the perfect combination of, on one hand, the museum's vast Van Gogh documentation and specialized documentalists and curatorial staff, and on the other hand the expertise in scholarly editing at the Huygens Institute. The editorial team was stationed at the Van Gogh Museum in Amsterdam.

An editorial board was inaugurated and a steering committee (directors and management team members of both institutions) guarded the financial parameters and decided on planning issues. It was everyone's ambition to do it properly once and for all, to aspire for 'the definitive edition' (even though everyone knows that in life only one thing is definitive and it is not a text edition). I am mentioning this because this shared dedication created a certain trust and latitude for the project to take its necessary course; to adjust the initial plans when this appeared necessary to maintain the high standards we had set for ourselves; and not least of all, to allow the editors more time than initially planned. It is not that we were free to do as we pleased; the point is that this shared sense of responsibility towards art-historians, art lovers and other readers all over the world was crucial to the eventual outcome and its success.

Aim

Our assignment was to prepare the manuscript for a scholarly edition of Vincent van Gogh's complete correspondence within five years. Initially it was expected to comprise about 12 to 14 volumes in printing, intended for an international readership and containing the edited texts of all the letters and a new English translation of them. The letters would be fully annotated and introduced by essays on general topics related to Van Gogh, his background, his letters and his correspondents. Needless to say, this bulk would be made accessible by a very cleverly designed index or multiple indices. Of course the little drawings in the letters, which were dubbed letter sketches, would be illustrated.

It is important to mention a particular secondary aim, which greatly influenced the editing process and thus the duration of the project. For both practical and conservational reasons, requests by scholars to view and study the original manuscripts at the Van Gogh Museum were seldom granted. One of the ideas behind the project was that it would produce a diplomatic transcription of the manuscripts which, in combination with black and white photographs, would make it unnecessary for scholars to see the original, sometimes very fragile letters. This service to scholars had serious consequences for the transcription method and for the way in which we presented these transcriptions; they constituted a kind of archival edition for internal use, in the form of digital files created with a conventional word processor, and available to visiting scholars.

Stages

The way we structured and planned the editing process followed a pattern that, I think, shows the obvious and natural order of consecutive stages. The basis for everything is the diplomatic transcription of the manuscripts in which as many characteristics as possible of Van Gogh's handwriting and writing process were documented.

Next comes the editing of the text, the result of which differs from the diplomatic text in that it is cleared of confusing and oftentimes unintended features and errors that may slow down or puzzle the reader. While we wanted to maintain the characteristics of Van Gogh's writing style, the urgency and dynamics of which made him sometimes sloppy at spelling and certainly at punctuation, it is the editor's duty to solve textual problems for the reader. And we wanted the edited text to be quotable without burdening those who quote a passage with a text that presents reading or interpretation problems.

This reading text, as we called it, is the basis for the next version, the translation. We had trial translations done that enabled us to write detailed guidelines for the translators, who had to work in a consistent and uniform way as much as possible. After a selection procedure we engaged a team of five professional translators and discussed their results in detail with the chief editor of translations. This was a costly project in its own right, and a very time-consuming one at that.

While translators were struggling with the pitfalls of Van Gogh's use of language we started research for the annotations and commentaries. We made use of well over twenty interns who did a lot of basic research.

When, after these various stages, we felt we grasped most of the key elements of Van Gogh's thinking and writing, we were able to write more elaborate introductions to general subjects that play a role in and behind the correspondence.

Two important lessons can be learned regarding such complex long term editing projects. The first one is that it is essential that, before definitively starting a new phase, an analysis be made of the material in order to investigate and try out what method works best in the long run. The results of each trial and possible options to proceed were always discussed with the editorial board.

The second lesson is that when many different people are involved, some shorter, some longer, the loss of consistency and coherence in the final result is a serious risk. It is therefore essential that the chosen method and all decisions and choices are documented in guidelines that must be applied consistently over the complete corpus and by all team members and collaborators.

Changes

I have become convinced that a disciplined and phased approach like I sketched is necessary for such a long term project to be successful. At the same time I have learned that it is impossible to foresee how such a project will evolve or how the outcome will compare to the original plan. Since each stage takes quite some time, say two to three years, one can in all reasonableness plan ahead only the first next stage. Depending on the environment in which the project is carried out, factors of different nature may influence its course. The Van Gogh Letters Project was carried out at the Van Gogh Museum. The editors became involved in several exhibition and publication projects, both initiated by the museum and by external partners; they were asked to take on other positions at the museum for a certain period of time; and they delivered many talks and lectures. These activities were one of the reasons why the project took fifteen years. In other words, it could have been done quicker. However it is my firm belief that the quality of the edition strongly benefited from these extra activities. They caused a lot of interaction and discussion with exactly those people for whom we actually were making this edition and it made us aware of the expectations and needs among colleagues and students, beside the fact that we ourselves were given a lot of useful research results in return. Also the extra activities created good publicity for the project.

There is one issue that greatly influenced the duration of the project that I have not mentioned yet. Of course digital technology developed so fast that around the turn of the century the dreaded question was raised: should we consider publishing on an electronic medium of any kind? We were not blind to what was happening in the publishing world. For a while we talked to a variety of commercial IT and web entrepreneurs who were emerging all over the market. They were eager to associate themselves with Van Gogh and the Van Gogh Museum because it would be great publicity for their companies. Our hesitation to join parties with them was caused by the fact that we encountered no true understanding for the nature of our work. We were in the process of producing 900 files with diplomatic transcripts, 900 files with edited texts, 900 files with translations of these texts, 900 files with notes, maybe 1500 illustrations of different kinds. Lots of pages with introductory

essays and other commentaries were in the making. They were not simply a load of files; they formed a complex structure of connecting layers and compartments, thematic networks, references, etc.. We believed that this was our contribution to Van Gogh scholarship and art history: not simply a compilation of files, certainly not a database – an easy way to irritate me is to call our edition a database. Our aim was to present and unfold the complex intellectual world that lies within Vincent van Gogh's extraordinary letters. It took us years to understand this complexity and to find a way of making it accessible and understandable. With the printed book in mind we had devised a system to help the reader in following and understanding the many possible ways to access the different versions of the letters, the thousands of notes and even more cross references, the wealth of documentary material (images, city maps, family trees, glossaries etc.), the commentaries and the apparatus. By simply dumping the files on the internet, to put it bluntly, the most valuable part of our knowledge and achievement would have been lost.

However, after a long and tedious process to find this solid academic press with whom we were on the verge of signing the contract, the publisher withdrew at the last moment, finding it too risky an investment. It was around 2003, and we were anticipating the finalizations of the manuscript part with texts, translations and annotations in 2006.

If the publisher's withdrawal was one sign of the times, there was another such sign that finally helped us to overcome the setback. Exactly what caused the publisher to withdraw, i.e. the increasing persuasiveness of the possibilities of electronic publishing, led the Huygens Institute to set up an IT department with the aim of becoming one of the frontrunners in digital editing. To build the Van Gogh edition would create opportunities for exploration and for gaining expertise. To us editors it was a relief that we would not be bullied by commercial guys but helped and guided by non-profit experts. We saw the advantage of presenting the full-fledged scholarly results, provided that the interface would be able to do even better what we had wanted to do in the books: to connect the different layers and levels of information, texts, images etc. The very fact that we already had thought that through was of great help to the development of the website, simply because we knew what we wanted.

We then decided that the very content of the edition would be extended: apart from the edited texts and their English translations we chose to publish the complete digital facsimiles and line-by-line transcripts of the text as well. In other words, Van Gogh's letters could be read in four different versions. And in a digital edition the number of illustrations is more a matter of choice than of money and if one can get a colour image, this can be used with no extra cost.

As a consequence, though, we had to mobilise interns and colleagues to assemble all this material. The digitization of the manuscripts was undertaken as a separate project, largely funded by Metamorfoze, the Netherlands' national program for the preservation of paper heritage. The publishing department of the museum helped to get the images together of all the artworks, both by Van Gogh and by the artists he mentioned. The Word-documents were converted into xml-files. These were all time-consuming things and they required a lot of discussion, choices and decisions among the editors, IT people and, at a certain point, the designers of the visual

presentation on screen. I am mentioning this to stress the fact that no matter how technical the issues are, no matter how many other specialists are needed to create, build and design the publication, it is the editors who must have the final say. It is their knowledge and vision that determines the scholarly value of the project. That said, the success of the publication, its accessibility, its transparency and user-friendliness depend to a high degree on the expertise of the collaborating IT people and designers.

Spin-off

As I explained earlier, we were all aware that the Van Gogh Letters Project was a one-off undertaking. It was hoped that the all-encompassing scholarly edition would lead to different types of publications to serve different segments of the readership. This worked out really well. A complete, illustrated and annotated edition was published in three languages at the same occasion as the launching of the website in 2009. (In 2016 it will be published in Chinese.) This is a luxury edition, with abridged notes. In the following years we, the editors, selected a more affordable anthology, with an extensive new introduction. This selection is now available in Italian, Dutch, Norwegian, French, Turkish and English editions; in the near future we anticipate the German, Japanese and Arab editions to come out. With this cheaper book we reach a much wider audience and the more literary oriented readers.

The Mondrian Edition Project

To the lessons learned I mentioned earlier, a few more can be added. First of all, interaction and collaboration with other scholars/projects takes time but can be very useful. Secondly, the interface should be the (digital) reflection of the editors' understanding of the edited source. And a scholarly edition can be the basis for derived publications; not the other way around.

These lessons were taken into account when the RKD – Netherlands Institute for Art History and Huygens ING decided to collaborate and initiate the Mondrian Edition Project. The Dutch artist Piet Mondrian has greatly influenced 20th century art and design and publications about his artworks fill many meters of bookshelves. Contrary to his image as a hermit he had a widespread international network and corresponded with many friends, artists, art dealers, collectors and curators. Very little of his correspondence has been published; his theoretical writings, very important to him as a means for spreading his idealistic views on art and society, and very influential to modern abstract painting, have been published in part and not in a critical edition. Mondrian's influence on modern art is at least as great as that of Van Gogh and a scholarly edition of his letters and theoretical writings has been a desideratum for many decades.

The Mondrian Edition project is again going to be a long term venture. In 2014-15 we conducted a one-year pilot project with the aim of investigating what approximately it may take in terms of money, time and content. We experimented with a small selection of 50 letters, and some theoretical texts. Peter Boot from

Huygens ING and DiXiT fellow Elena Spadini created an xml schema for the letters and we did a lot of calculating and estimating. This led us to think that we will need approximately 12 years to finish.

Not the same

Despite the similarities with the Van Gogh situation, it will not be a matter of copying the Van Gogh approach or format, nor even the methods applied. An important difference is the almost double number of letters (namely 1600 instead of 900), a more diverse group of correspondents (namely 125 instead of fewer than 25) and of course the theoretical writings we want to include. These writings have survived in the form of publications, manuscripts, typescripts, notes etc. and to reconstruct their writing and publication process will need a different editorial approach than the letters. However, Mondrian's theoretical thinking and writing, and the publication of the resulting texts are so intertwined with large parts of his correspondence that in our opinion it is appropriate to present these different types of texts together. How to combine them in an organic and transparent way is one of the new challenges of the Mondrian Edition Project.

Another difference in comparison to the Van Gogh Letters Project is the fact that we intend to include the wealth of documentation that is kept by the RKD and some other institutions and that is cited or requested frequently by scholars. This regards Mondrian's personal documents, his horoscope and portrait photographs but also very important photographs of his various studios, installation pictures, exhibition catalogues etc.

The RKD publishes online several databases containing all kinds of art-historical and archival information about artists, artworks, publications, auctions, exhibitions etc. Instead of copying in that sort of information in our edition, particularly in the annotations, we intend to harvest these data through links to the RKD database files; the same could be done with databases from other archives, libraries, museums and the like if they are willing to cooperate.

The next innovative addition is the complete catalogue *raisonné* of Mondrian's paintings and drawings. It originally was published in print in 1998 and it is an essential reference book for all who are interested in or doing research on Mondrian. The updated, digital version is being prepared in conjunction with the edition project and, as said, will be part of... what shall we call it? Is it still an edition? The new digital opportunities challenge us to be creative and to push the limits of the traditional concept of a text edition. Of course we could simply call it a website but from a theoretical point of view we as scholars are challenged to engage in some methodical and theoretical, if not paradigmatic re-orientation.

Since the launch of the Van Gogh edition a lot of progress has been made in digital publishing and digital humanities. Also, in recent years a range of questions have surfaced – about crowd sourcing, new tools, sustainability, etc. – that need to be addressed in the near future if we want to make sure that our work remains relevant and accessible to future generations. The DiXiT fellows are the new generation whose task it is to give directions for the future of digital humanities. I thank them in advance for their contribution to the Mondrian project.

Digital editing: valorisation and diverse audiences

Aodhán Kelly¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

The perceived potential to reach broader publics has been an oft-discussed topic since the earliest adoption of digital technologies for the publication of scholarly editions. A dichotomous concept of 'audience' has become prevalent in the field, divided into 'scholarly' and 'non-scholarly'. John Lavagnino (2009), for example, described it as 'the problem of two audiences' while Edward Vanhoutte (2009) has proposed an integrated dual-model of maximal and minimal editions, one aimed at each audience. This dichotomy typically renders that there are two potential outputs: the 'scholarly edition' aimed at scholars, and the 'reading edition' for a general public. But what happens beyond this, what other ways does digital editing attempt to valorise scholarly knowledge in society at large?

In this paper I will take the stance that there is, rather, a diverse spectrum of publics who possess overlapping layers of interests, competences and capacities. Not only can audiences be extremely diverse but so can the potential means that could be employed to engage them. What forms of engagement do we value and how? Spin-off publications aimed at general audiences often are considered to be a form of 'outreach', a side-project to the scholarly edition itself but is that a healthy perspective to take on scholarship? Are we limiting our efforts geographically or should there be attempts to engage more globally diverse audiences? This paper will offer a state of the field in efforts to valorise scholarly knowledge with more global and diverse audiences, and attempt to suggest some potential avenues for further exploration.

My research role with the DiXiT Network is to investigate the dissemination of digital editions by looking at ways in which we can communicate textual scholarship with broad audiences, both specialist and nonspecialist alike. Primarily, I am trying

¹ aodhan.kelly@uantwerpen.be.

to look at digital editions through the lens or perspective of the user. Looking at perspectives of Engagement, Discoverability, Usability and Accessibility for different user types. Typically speaking when I describe my research area to other scholars in the field it is met with some welcoming positivity as they are trying to figure out ways to valorize their research.

What is valorisation? Knowledge valorisation refers to the utilisation of scientific knowledge in practice. Examples include developing a product or a medicine, or applying scientific knowledge to a system or process. It is something which has come originally more from Science, Technology and Medicine fields but increasingly is applied to humanities research. While valorisation is not something which I initially thought I was trying to address in my research it is clearly something which is on the tip of peoples' tongues, a genuine concern that is surely the sign of an increasing impact agenda within humanities funding structures. Valorisation in the humanities is also akin to the idea of 'impact' which Simon Tanner defines as 'the measurable outcomes arising from the existence of a digital resource that demonstrate a change in the life or life opportunities of the community' (Tanner 2012, 4).

There are a variety of perspectives on this from various funding and review bodies such as the AHRC and REF in the UK. But instead of discussing impact measurement metrics today I intend to highlight some digital approaches to social valorisation and social impact that primarily address the non-scholarly audiences. In order that we that might be able to foster ideas for digital editions projects about how to allow our research to have a positive impact on broader society. Then, almost by proxy, addressing some of these funding requirements. In order to do this I will look at a number of case studies of digital projects as well as looking at some other potential approaches for consideration. While there are many different ways we could look at this question I would like to focus on three broad areas for today's discussion: Social and Public Approaches; Addressing Digital Divides; Physical spaces in GLAMs (Galleries, Libraries, Archives and Museums). These are potential avenues of exploration that I have identified through my research into dissemination methods.

Social and Public Approaches

Perhaps it stands to reason that one of the best ways to create social value is to take social approaches in creating our digital resources, or in other words, to engage in public humanities. Many of you will have heard much about the *Transcribe Bentham* project which has harnessed crowdsourcing for the purposes of transcription. Given the timing of this paper I would rather talk about the *Letters of 1916* project at Maynooth being only a few weeks before the centenary of 1916 Easter Rising in Dublin. The project is the first public humanities project and has over 1300 volunteers working on the transcription of digitised letter collections from the year 1916 in Ireland. The most interesting aspect of this project, from a social value creation perspective, is indeed the public aspect. The history of the 1916 Rising, like many similar events in other countries, has developed a sort of mythic status and still remains an emotionally charged topic. It has been used and

abused by various parties in Ireland during the period of national identity building post-independence and throughout the 20th century and sometimes forced into single narratives, which is only beginning to be countered in recent years. This type of public humanities project has the potential to deconstruct the myths and redress this type of single narrative. It involves 'ordinary people' transcribing the letters of 'ordinary people'. This process results in a democratisation of knowledge and that truly can be considered to be a form of social impact. When addressing subjects in which the events lie distinctly within living memory it could have even more value. The *Genocide Archive of Rwanda* for example has been collecting materials and testimonies since shortly after the tragic events of 1994 in which approximately 20% of its population died. They hope that through education on these events it might help to prevent future genocidal atrocities.

The *Readers' Thoreau* project is an excellent example of a social approach to digital editing but with more of a pedagogical focus. It combines a digital reading edition of the *Walden* text with an online community of engaged users who can avail of a collaborative annotation functionality. This format creates an interesting discourse, whereby annotations from well known scholars sit alongside comments from current users, gathering different generations of scholarship together in one place, as if in conversation with each other. The editors see this as a tool for democratic practice which can build community and empower students. The importance of annotation is recognised as one of the scholarly primitives laid out by John Unsworth (2000). In my mind this project represents a real and original contribution to the field of digital editing and its methods of user engagement and pedagogical experimentation. It is a format I believe could be replicated easily, although I do not think it is available on GitHub.

Addressing Digital Divides

This leads me into the next area which is the discussion of digital divides. This term typically is used in relation to the divide in digital access to information between wealthy nations and developing nations. But it can also be a divide in access for those with visual impairments or those users with lower levels of literacy. It would perhaps be fairer to talk about this across five categories of divide: literacies, socio-economic, technological, disabilities and linguistic. While we are usually in the business of talking about purely textual matters I do believe that modes of communication other than reading are also important as demonstrated by Claire Clivaz.²

Erik Champion in his work on critical gaming argues that while text-based research is highly prevalent in DH that non-textual approaches and media are important, relevant and can augment rather than replace text. He claims that: 'A concern or predilection with text-based material is obstructing us from communicating with a wider audience. Multimedia, visualisations and sensory

2 See Claire Clivaz, 'Multimodal Literacies and Continuous Data Publishing: *une question de rythme*'. Opening keynote given at *Academia, Cultural Heritage, Society DiXiT Convention*, Cologne, March 14-18, 2016, this volume.

interfaces can communicate across a wider swathe of the world's population. And although literacy is increasing, technology is creating a fundamental divide between those who can read and those who cannot' (Champion 2015, 10). He also points out here that according to UNESCO today that the rough global percentage of literacy today has risen to 84% from 76% in 1990. That still means that 774 million people are functionally illiterate and about 160 million of these are in highly developed OECD nations. The *eTalks* platform demonstrated by Claire Clivaz is a good example of how to address different types of literacies simultaneously by linking together text with sounds and images. In her talk yesterday she also made a point that history usually cannot tell the story of the losers, but maybe through this multimodal approach it might be possible to do so. The Rwanda digital archive mentioned earlier is that it is truly multimodal, it contains documents, publications, video recordings, audio recordings, objects, photographs, interactive maps, transcriptions and commenting features and perspectives from both victims as well as perpetrators. This might be a good example of a digital resource that addresses Clivaz' concerns in this area.

Regarding the economic aspect of the digital divide it is worth pointing out that the *Rwanda Archive* was established and funded by a British NGO the Aegis Trust. Many other cultural heritage digitisation programmes in underdeveloped nations and conflict zones such as South Sudan and Syria depend entirely on external support. It is worth remembering that the very existence of these types of projects depends on the support of academics like the people in this room. Initiatives exist to make digital journal access free or affordable for developing world economies in STEM fields but I do not know if similar initiatives exist in the humanities nor how to approach this from a digital editing perspective.

Aside from addressing economic divides affecting access to textual cultural heritage there is also the issue of the technological gap. This is the digital divide in the most common sense of the expression. One of the biggest challenges regarding access to cultural heritage information (or rather access to information in general in these regions) is the reality of lower wifi speeds and the reality of lower standards of digital hardware available too. The Global Outlook Digital Humanities special interest group is beginning to explore this area, particularly with a focus on the concept of minimal computing. Which is using techniques like static website generation, strategies for data transfer, low cost and DIY hardware in order to make digital humanities resources available to those with technological constraints.

The remaining digital divides are those arising from disability and due to linguistic divisions. How are we to make our resources accessible to those with visual and hearing impairments? If we are to truly achieve multimodal communication of our textual cultural heritage then we need to consider this area. Many of you will already be familiar with some basic W3C accessibility guidelines such as adding Alt text descriptions to images to make them machine readable and there is also Text-to-Speech (TTS) technology that allows text from websites to be read aloud. Decisions about which languages we make our resources available through naturally are going to affect the composition of the audiences we can reach, but how do we decide? Needless to say, there is a great need for increased discussion of these topics in our field.

When we talk about social impact I would implore that we should try to: think globally and act locally. This is an expression that typically is associated with environmentalism. Although in this case it usually refers to ensuring that local activities should seek to limit any negative impact on the more global environment. In the case of our cultural and scholarly work it would be more along the lines of seeing how we can make our local activities have a positive impact on the global.

Physical spaces in GLAMs

The third and final area I want to bring into discussion is what we could do with digital edition content in public places and physical spaces like in GLAM institutions exhibitions. The importance of public museums and other similar institutions for access and exposure to cultural heritage is immense, even more so for people from less privileged economic backgrounds. Most people who walk through the doors of a GLAM institution are there in order to gain access and exposure to culture and heritage thereby making them an ideal environment in which to try to make a social impact with its relatively captive audience.

During a visit to *Archives+* in the Manchester Central Library I was inspired into the idea of creating digital touchscreen exhibitions in GLAM institutions as a form of public engagement for digital editing projects. *Archives+* is a truly multimodal and multimedia space, open, welcoming and engaging and is attracting a real public interest with its use of digital exhibitions. This is the one of the three areas which I intend to research with the DiXiT members based at Antwerp, we are currently in the initial planning phase for such an interface in a literary museum in the city. You need to adjust your expectation for what you can communicate on these types of devices compared to a web based digital edition. I conducted some user research last year with Elena Pierazzo at King's College London on the potential use of tablet devices for the dissemination of digital editions. We found that tablet users were less interested in accessing a full-on digital scholarly edition with critical apparatus et al in this environment but could see a real value in their usage for reading editions, learning, and public engagement (Kelly 2015, 137). These fixed digital touchscreen exhibitions in GLAMs are then probably most suited to purposes of learning and public engagement in these spaces. I am interested to discover whether this kind of partnership between memory institution and a scholarly editing project could be a mutually beneficial exercise in the dissemination and valorisation of textual heritage knowledge.

Conclusion

All three of the areas I have raised in this paper still need a fuller investigation from within our field but I hope by raising this today I can stimulate some discussion of how this might be done. Perhaps by engaging in social and public approaches to editing, by thinking more globally about our idea of social impact and through a reconsideration of some of the physical spaces in which society comes into contact with textual cultural heritage maybe we can achieve and demonstrate real social value even beyond what digital scholarship has already achieved in this arena so far.

References

- Archives+*. Accessed March 3, 2017. <http://www.archivesplus.org/>.
- Champion, Erik. 2015. *Critical Gaming: Interactive History and Virtual Heritage*. London: Ashgate.
- Genocide Archive of Rwanda*. Accessed March 3, 2017. <http://www.genocidearchive-rwanda.org.rw/>.
- Global Outlook: Digital Humanities*. Accessed March 3, 2017. <http://www.globaloutlookdh.org/>.
- Kelly, Aodhán. 2015. 'Tablet computers for the dissemination of digital scholarly editions'. *Manuscripta* 28: 123-140.
- Lavagnino, John. 2009. 'Access'. In *Computing the edition*, edited by Julia Flanders, Peter Shillingsburg and Fred Unwalla. Thematic Issue of LLC. *The Journal of Digital Scholarship in the Humanities* 24 (1): 63-76.
- Letters of 1916*. Accessed March 3, 2017. <http://dh.tcd.ie/letters1916/>.
- Readers' Thoreau*. Accessed March 3, 2017. <http://commons.digitalthoreau.org/>.
- Tanner, Simon. 2012. *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*, King's College London, October. Accessed March 3, 2017. www.kdcs.kcl.ac.uk/innovation/impact.html.
- The eTalks*. Accessed March 3, 2017. <http://etalk.vital-it.ch/>.
- Unsworth, John. 2000. 'Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?' *Paper presented at the symposium on Humanities Computing: Formal Methods, Experimental Practice (King's College, London, 13 May)*. Accessed March 3, 2017. <http://web.archive.org/web/20160604163440/http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html>.
- Vanhoutte, Edward. 2009. 'Every Reader his own Bibliographer – An Absurdity?' In *Text Editing, Print and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland. Aldershot: Ashgate, 99-110.

Social responsibilities in digital editing – DiXiT panel

Editing and society: cultural considerations for construction, dissemination and preservation of editions

Aodhán Kelly¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

This presentation intends to initiate discussion during this panel regarding the social role of our scholarly editing community in society at large. This is slightly off-topic from the general theme of this conference on technologies, tools and standards but, nonetheless, something that I feel is important to bring up on a panel focusing on 'Editing and Society'.

I was inspired into this line of discussion when I recently had the unusual privilege of visiting what is probably the newest national archive in the world, situated in an extremely and tragically war-torn country. I prefer to keep the country anonymous in this discussion so as to avoid any chance of further endangering the archive. The foundation of the archive in question was the product of the hard work of a small group of scholars who assembled materials from around the country in difficult circumstances, catalogued them using recognized standards and have stored them in donated archival boxes in a cool dry building. Operating without a single shelf in the entire facility, they still manage to remain relatively organised and grant access to the small number of researchers who manage to visit.

It is operated with pride and enthusiasm but without a formally approved mandate from its very volatile government. Housing almost the entirety of its written historical record for the 19th and 20th centuries in one place, they are

¹ aodhan.kelly@uantwerpen.be.

acutely aware that if the government should decide that the documents housed in the archive might threaten their power in any way, that the entire collection could be razed at any time. An external cultural heritage NGO is assisting them in digitizing their collections, in order of priority set out by one of the scholars, and should this worst-case scenario occur there is at least some hope of the digital record surviving externally. Needless to say this visit stimulated a real interest for me.

Moving on to another region of the world, we all have been reading about the ongoing destruction of important cultural heritage in Syria and Iraq and it has begun to generate discussion in the media throughout the world particularly regarding the destruction of the site at Palmyra. There are global organisations working on emergency preservation such as UNESCO or Global Heritage Fund. The principle focuses of these kinds of organization relate to built cultural heritage as well as intangible cultural heritage (music etc.). However, there is considerably less focus and funding for preservation of the documentary record. There are exceptions such as the British Libraries endangered archives programme. And specifically in Syria the Hill Museum & Manuscript Library at St. John's University from Minnesota have been running a rather successful digitization programme since 2011.

Is it a scholarly community responsibility to get involved in preservation?

I do not really have a clear answer to this question, as it is a topic in which I am perhaps not yet well informed enough, but I do believe strongly that it should be discussed or at least acknowledged at an event like this. This is a topic discussed far more often among GLAM (Galleries, Libraries, Archives, Museums) circles than the scholarly editing community as such. However, I do remember from a previous DiXiT meeting in Sweden earlier this year that when we were asked what is at the root of what we do as members of this particular academic field, one answer was that we should take on the role of 'custodians of cultural heritage', or something to that effect. Should scholarly groups such as ours gathered here today, as experts in textual cultural heritage objects and digital methodologies try to engage more with these challenges, and what roles could we take?

A more global outlook

To achieve that line of thinking probably requires us to think a little bit more globally, sometimes outside of our normal spheres of engagement. In the DH community more generally there is a slowly increasing interest in this kind of attitude, such as with the ADHO special interest group Global Outlook::Digital Humanities (GO::DH). Their aim is to help break down barriers to communication and collaboration in research and study – between countries of differing economic status – in digital arts, humanities and cultural heritage. 'Minimal Computing' is one of their research areas for example. It focuses on computing done under some set of significant constraints of hardware, software, education, network capacity, power, etc. It acknowledges the reality of conditions in many other parts of the

world. This is what commonly is referred to as the ‘digital divide’. Simon Tanner argues that ‘whilst we can hope for the digital divide to be eradicated, it might be more reasonable to expect information technology to address some of the worst information, education and cultural resource inequalities rather than solve all of them’ (Tanner 2005: 27). DH in general seems to be making headway in this area, but should we be talking about it more often in the scholarly editing community?

Should we also treat dissemination as a scholarly responsibility?

I think it is not that difficult for us to think about preservation of cultural heritage as something with which we feel a certain amount of social responsibility but would we be able to apply the same feeling of obligation to the dissemination of knowledge and scholarship to society at large. My research role in the DiXiT network is to investigate dissemination of digital editions. A lot of my focus is on figuring out how to convey important aspects of this type of scholarship to broader (often non-scholarly) audiences often using potential spin-off publications that stand independent from the edition itself.

But who are our audiences?

Making false assumptions about the composition of our audiences can be hard to avoid unless we are consciously making an effort to assess it, such as through the use of analytics software for example. One interesting case I heard at the MLA 2015 conference in Vancouver came from Neil Freistat who explained that during the first 24 hours after the *Shelley-Godwin Archive* was launched the site received 60,000 visits, which came disproportionately from Latin America and Eastern Europe. This really made them re-think about what audiences that they might reach. It highlights that while the majority of digital editing projects are taking place in Western Europe and North America, the end users of these editions can be far more diverse.

So maybe we should prepare for the possibility of diverse users. Who exactly should we try to engage with? We cannot realistically engage with everyone. Not every piece of scholarship can gather broad appeal, but many can and many have something to communicate even if they need to do it by other means than the edition itself. Elena Pierazzo discusses different purposes for reading works of literature ‘one can read to enjoy the content or for the purpose of studying some particular feature of the text; some may also be interested in the transmission history of the text’ (Pierazzo 2015: 153). She then points out that a digital environment can support (and enhance) all of these types of readings, but not necessarily within the same edition nor even within the same environment’ (Pierazzo 2015: *ibid*). The *CantApp* created by Peter Robinson and Barbara Bordalejo is a good example of what I mean by a spin-off publication that is not the edition itself, but is derived from its scholarship and has the potential to engage a different audience and generate popular interest in a difficult but important text. The Jane Austen manuscripts project which received a lot of press after its launch due to

die-hard Austen fans discovering that their favourite author's writing style was far from fluent, with many typos and vast amounts of corrections, shattered popular myths about her manner of writing. Generating this non-scholarly interest alone is a fantastic achievement in my opinion.

Dissemination activities and concerns

In my opinion dissemination involves a wide range of concerns and activities that can be divided into four broad categories: engagement, discoverability, usability and accessibility.

Engagement

There are numerous potential ways to engage diverse users with scholarly texts such as the creation of reading editions, pedagogical exercises, gamification, social editing or crowdsourcing activities, social media or utilising physical exhibition spaces like museums, galleries and classrooms.

Discoverability

How do we make our publications discoverable for the appropriate audiences? A survey I conducted in 2014 revealed that even among scholarly users the most common ways they discovered editions was through academic citations and word of mouth (2014: 131). This is despite the fact that there are discovery resources available such as digital catalogues those by Patrick Sahle or Greta Franzini as well as subject-specific resources. Other discoverability concerns include search engine optimisation (SEO), the application of appropriate metadata standards and compatibility with larger knowledge consortiums and cultural data aggregators such as Europeana for example.

Usability

A user-centric perspective is vital to ensuring successful (and enjoyable) usage of digital outputs and thus, naturally, user-centred design approaches which often may require conducting various forms of user studies. Implementing user analytics software in order to better understand the user paths through a digital publication as well as its possible strengths and weaknesses can provide invaluable insights and direction for improvements in design.

Accessibility

Ensuring that a publication is accessible to the appropriate users is yet another many-headed hydra. This can involve a delicate balancing game between adhering to Open Access principles while navigating various copyright concerns. As mentioned earlier it is also important to have an awareness of how to make scholarship available to those with lower technological capacities or equally to consider users with disabilities. Accessibility of open-source data for remix and re-use in further scholarship also falls under this category.

These are the manifold concerns of which I believe that the dissemination process is or should be comprised. To be effective in this it is crucial to create and implement actual dissemination strategies and to do so as early in the editorial process as possible. By treating questions of dissemination as a mere afterthought there is a risk of developing an edition that excludes or limits many of the potential users. These are the activities that need to happen if we want our scholarship to reach our intended audiences and diverse ones too.

Conclusion

In conclusion I actually have two separate questions for this panel discussion. Firstly, should we embrace more of the social responsibility of engaging in the preservation of cultural heritage using the skills and means available to us as a community? Secondly, could we begin to see the dissemination of our scholarship to society at large not just as something necessary to meet funding and professional requirements but more as another form of social responsibility?

References

- British Library Endangered Archive Programme <http://eap.bl.uk/>.
CantApp. <http://www.appbrain.com/app/cantapptest/com.sdeditions.CantAppTest>.
Europeana. <http://europeana.eu/>.
Franzini, Greta. *A Catalogue of Scholarly Digital Editions*. <https://sites.google.com/site/digitaleds/>.
Global Outlook::Digital Humanities (GO::DH). <http://www.globaloutlookdh.org/>.
Global Heritage Fund. <http://globalheritagefund.org/>.
Hill Museum and Manuscript Library. <http://www.hmml.org/>.
Jane Austen Fiction Manuscripts. <http://www.janeausten.ac.uk/>.
Kelly, Aodhán. 'Tablets for the dissemination of digital editions.' *Manuscripta* 28 (2015): 123-140.
Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate Publishing, 2015.
Sahle, Patrick. *A Catalog of Digital Scholarly Editions*. <http://www.digitale-edition.de/>.
Shelley-Godwin-Archive. <http://shelleygodwinarchive.org/>.
Tanner, Simon. *Digital Libraries and Culture: A Report for UNESCO*. King's Digital Consultancy Services, 2005.
UNESCO World Heritage Centre. <http://whc.unesco.org/>.

Documenting the digital edition on film¹

*Merisa Martinez*²

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Introduction

Film is used for ethnographic data collection in the social sciences (Banks 1995) but less so as a method of disseminating research (Baptiste 2015) or documenting the creation of academic work in the arts and humanities (Bell 2006). The first two decades of the 21st century saw an explosion of online learning environments and eventually MOOCs, which often disseminate knowledge in the form of short digital films. This medium reaches wide audiences through easily accessible, sharable and usable interfaces like YouTube and Vimeo. Digital films are modular, reusable, redistributable, and with a growing market of low-cost equipment, can be scaled up or down to fit many different project budgets and levels of expertise. Given this potential, the short academic film can be viewed as a complimentary method to document the creation of digital editions and explain their technical and intellectual content.

Documentation as a necessary part of academic output

In the hard sciences, replicating the results of a proof or experiment is a way of validating those results. Replication necessitates an accurate description of the steps taken to form the results in question. This is made geographically complex in the

1 This brief presentation was given as part of the panel 'Editing and Society: Cultural considerations for construction, dissemination and preservation of editions.' Rather than presenting research results on a finished experiment, it is intended to provoke conversation about one aspect of a forthcoming project on scholarly filmmaking in the digital humanities.

2 merisa.martinez@hb.se.

digital scholarly editing community, where collaborative networks spread out over countries or even continents is the norm, and the considerable task of documenting meetings, building websites, metadata and mark-up choices may be distributed unevenly throughout the (often multi-year) lifecycle of a digital edition. And though digital scholars may apply previously developed tools, theories or editing styles to their projects, innovation rather than replication is the goal. Nevertheless, there has been a significant uptick in digital project collaborators documenting their editorial practice through conference presentations and in user guides on edition websites. Documentation can significantly affect how a digital resource is used and reused (Warwick *et al.* 2009). If users have difficulty understanding where to start with a digital resource, they may simply choose to exit. Conversely, creating textual and audio-visual guides with detailed descriptions of the technical and academic content therein can be a good way of increasing users' time on the website, and could potentially increase engagement with its materials.

Understanding the technical infrastructure and theoretical underpinnings that lead to the creation of a digital edition also can engender a sense of appreciation for its value as a scholarly/art object (Bell 2006). Within the scholarly editing community, text has an understandably significant hegemony, but particularly in the case of digital editions, visuals are arguably as important as the text. What 'counts' as a digital edition increasingly has concerned scholarly editors. Indeed, Elena Pierazzo argues that only a combination of digital 'methods of production' and their research outputs can qualify a work as a *digital* edition (Pierazzo 2015). With the introduction of digital methods to the creation of scholarly editions, editors either have gained technical skills in user-interface and graphic design, collaborated with technical staff who take on these tasks, or a combination of the two. This adaptation to new methods and collaborations shows that it is possible to address issues that come up in a specifically digital work environment. It also indicates that it is possible to address the needs of diverse digital user groups with the creation of further audio-visual aids.

Film as documentation

One way of using film to document the digital edition is by creating a visual user-guide as an ancillary video for the edition. This could be as simple as a screencast showing a narrator clicking on different links in the edition, with a voice-over and subtitles explaining its mark-up style, or its modules, where to find the critical apparatus and how to download the XML files (to name just a few options). Films could be broken down into different competency levels or familiarity with digital editions, *e.g.* 'beginner', 'intermediate', and 'advanced'. Opportunities abound for creativity in this medium. Further, creating a user-guide is a good exercise for scholarly editors and others who work on digital editions to think through how their editions will be used and *how* to effectively communicate the options for use and reuse of data. Film is a medium well-suited to couple with traditional academic output like research papers, written user-guides and conference presentations about projects. Diverse user groups receive something that is visually demonstrative, but

also allows – through easy interfaces – for users to start, stop, pause, go back, repeat, increase or lower volume, increase or decrease window size, and perform other simple functions which customize the viewing experience.

Perceived Problems

One perceived problem is simply where to start if a scholarly editor has no experience in filmmaking. Luckily, there are many written guides and short films on how to make short films. There are also workshops (many of them free) and subscription-based websites and MOOCs which provide simple instruction on how to use equipment, how to storyboard and edit, and how to compress and code short films. Perhaps the greatest reason to choose this medium is that it is another opportunity to collaborate: many university and public libraries create short films about their online and in-house exhibits in order to gain public awareness. If they have an in-house production team, it is possible that the team can provide supplies, staff, and advice to a digital editing project. There is no one way to do this – film editing and output is similar to the TEI in that there are standards but they can be employed in different ways for different people and different audiences.

Another perceived problem is expense. Getting started with digital equipment can be overwhelming, and with the added cost of some top-of-the-line equipment, the cost-benefit analysis can immediately outweigh any desire to get started in this medium. Luckily the market for affordable video-capable DSLR cameras and lenses has expanded in the last ten years, and there is a growing market of refurbished equipment as well. There are also many companies (some that ship to several countries) which rent out digital equipment. And, if a project is connected to a library or university, that institution may have its own equipment that can be rented by university-affiliated employees. Most DSLR enthusiasts start out with borrowed or rented equipment to get a ‘feel’ for this medium, so cost can be kept relatively low for first-time users.

A third perceived problem is time. Film is its own form of storytelling, and, even in short videos this can require a lot of time and energy, resources which may be in short supply. Using collaborative networks and thinking carefully about how to construct a sufficient user-guide while the edition is being built is a good way to offset time cost, as the work of storyboarding can be done in parallel with the work of building the actual edition.

A fourth perceived problem is peer-review. Digital project collaborators already face significant concern about how to properly assess the scholarly merit of their work. Yet it is also possible that by opening up this work to include films, one opens up the possibility of constructive review and assessments from new media scholars, film studies scholars, and others who have expertise in visual storytelling. And this type of film is not seen as a replacement for scholarly output, but rather as a support of it. There is still space for ‘traditional’ peer-review in the form of the written work of the edition, and through well-respected avenues of publication and dissemination such as conferences, academic journals, and monographs.

Conclusion

Taking time to document our processes, particularly with regard to collaborative efforts, can arguably be described as a cornerstone of digital humanities research. The current system of project-based output in the digital humanities could benefit from diverse methods of visual storytelling to aid a diverse group of users. This may prove a challenge to scholarly editors without the benefit of significant training in filmmaking, but the cultural heritage sector has proven expertise in this medium, and as such is an area where collaborations can be made. The viral nature of digital video through social media sharing can benefit digital projects as a way of grabbing attention and broadening audience interaction. In order to sustain attention, digital academic films must be focused on effective digital storytelling, and on the users. There are lots of different types of short films – documentaries, trailers, interviews, videos – that explore the materiality of a specific text or time period covered in the edition, and explanations of methodology; all of which can provide interesting ancillary materials to digital editions.

References

- Banks, Marcus. 1995. 'Visual Research Methods.' *Social Research Update*. (11). <http://sru.soc.surey.ac.uk/SRU11/SRU11.html>.
- Baptiste, April Karen. 2015. 'Can a research film be considered a stand-alone academic publication? An assessment of the film Climate Change, Voices of the Vulnerable: The Fisher's Plight.' *Area*. DOI: 10. 1111/area. 12194.
- Bell, Desmond. 2006. 'Creative film and media practice as research: In pursuit of that obscure object of knowledge.' *Journal of Media Practice* 7(2): 85-100. DOI:http://dx.doi.org/10.1386/jmpr.7.2.85_1.
- DeFillippi, Robert J and Michael B. Arthur. 1998. 'Paradox in Project-Based Enterprise: The Case of Film Making.' *California Management Review* 40(2): 125-139.
- Henley, Paul. 1998. 'Film-making and Ethnographic Research.' In *Image-Based Research: A Sourcebook for Qualitative Researchers*, edited by Jan Posser, 42-59. London, UK: Falmer Press.
- Lundén, Tomas and Karin Sundén. 2015. 'Art as Academic Output: Quality Assessment and Open Access publishing of Artistic Works at the University of Gothenburg.' *Art Libraries Journal* 40(4): 23-32.
- Parr, Hester. 2007. 'Collaborative film-making as process, method and text in mental health research.' *Cultural Geographies* 14: 114-138. DOI: 10. 1177/147447400772822.
- Pauwels, Luc. 2010. 'Visual Sociology Reframed: An Analytical Synthesis and Discussion of Visual Methods in Social and Cultural Research.' *Sociological Methods & Research* 38(4): 545-581. DOI: 10.177/0049124110366233.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Surrey, England and Burlington, Vermont: Ashgate Publishing.
- Thieme, Susan. 2012. "Action": Publishing Research Results in Film.' Forum: Qualitative Social Research. 13(1) <http://nbn-resolving.de/urn:nbn:be:0114-fqs1201316>.
- Warwick, Claire, Isabel Galina, Jon Rimmer, Melissa Terras, Ann Blandford, Jeremy Gowand George Buchanan. 2009. 'Documentation and the users of digital resources in the humanities.' *Journal of Documentation* 65(3): 33-57.

Towards a definition of 'the social' in knowledge work

Daniel Powell¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Culture and knowledge have always been social, produced and consumed within networks. The scholarly landscape is changing rapidly; not only is collaborative authorship becoming (slowly) more widespread, but innovative publication practices for academic work quickly are becoming mainstream. Alongside peer-reviewed journal articles, academic blogging is everyday practice for those working in many fields. At in-person conferences and symposia, there is often now a Twitter-based back channel posing questions, linking to materials, or carrying on discussion. Conference panels are put together on Facebook, and an institutional repository or Academia.edu page might matter more than a formal pedigree. The modern scholar lives in a deeply interconnected world of information, and the manifold connections between those bits of information, and between those bits and people are, inexorably, shifting long-understood processes of academic work. Knowledge work now is, or can be, transparently social, outwardly iterative, and incredibly fast-moving. My research attempts to grapple with the making of culture and of knowledge in the 21st century, and this brief piece is an attempt to begin thinking through 'the social' as a concept in contemporary knowledge work within the academy.

To bracket 'the social' as an adjective describing knowledge work is a problematic opening gesture, but one that nonetheless carries a certain force precisely because it highlights the antisocial nature of much academic work in the humanities throughout the last century. In *Keywords*, Raymond Williams sketches out a trajectory of society as a linguistic and cultural concept in modernity. By way of definition, he writes that '(s)ociety is now clear in two main senses: as our

¹ danieljamespowell@gmail.com.

most general term for the body of institutions and relationships within which a relatively large group of people live; and as our most abstract term for the condition in which such institutions and relationships are formed.² In his view, and backed up by eclectically selected samples throughout the last approximately 600 years, the development of ‘society’ is largely a history of movement from the Latin sense of ‘ally’ or ‘confederate’ through to ‘company’ or ‘companionship’ to the definition cited above. Along the way, Williams notes that the tension between the ‘general and abstract’ vs. the ‘active and immediate’ senses was always present, but that by the 19th century, the latter sense had been abandoned almost entirely in favour of the former. The idea of ‘the social’ followed a similar trajectory, moving steadily from a sense of genial association to objective generalizability. Thus while in up to the 17th century ‘social’ might appear as a synonym for ‘civil’ – as in the Social War of Rome against its longstanding allies in the 1st century BC – by the 19th century, ‘society’ as such had become externalised to such an extent that it became possible to speak of social reformers, social diseases, social geographies, social status, and so on – and in fact many of these terms and those like them formed during the emergence of society as an objective, outside construct during the 100 year period of 1810-1910.³

Many of the ways that we think about knowledge production, about remediation in an age of digital reproducibility, and about digital scholarly editing have been influenced deeply by the multivalent nature of society and the social as terms and constructs. For example, Elena Pierazzo’s recent book *Digital Scholarly Editing* summarises social editing in essentially two contexts: the first is the rise of social media connectivity and Web 2.0, and the second is developments in discourses of social textuality.⁴ The first set of ideas is obvious from the way many of us live our lives (there are nearly 1.5 billion monthly users of Facebook, for instance), and the second comes to us through Donald McKenzie and Jerome McGann.

I actually want to argue for a conception of social knowledge that is connected to both social media connectivity and to the analytical categories scholars might bring to bear, in the manner of McKenzie or Robert Darnton, on the inscription-bearing artefacts of knowledge work in the humanities. This leads to two claims: first, scholars must accept and take seriously the McKenzian view that all texts are constructed socially in both their production and interpretation. If applied to academic work, then it is difficult to claim that the outcomes produced in the academy – from scholarly editions to journal articles to white papers to undergraduate essays – are also not texts of some sort, subject to the same analytical categorization as McKenzie and others apply to historically distant works. For McKenzie, social production easily could be seen in the printing shops of early modern England, where the lines between authors, editors, publishers, and printers became hopelessly muddled. This is a way of understanding the production of textual artefacts. On the other hand, the McKenzian reading of land as text in pre- and post-treaty New Zealand is illustrative of how textual

2 Williams 1983: 291.

3 Ibid., 291-295.

4 Pierazzo 2015: 18-25.

meaning is intricately bound up in social norms, expectations, technologies, and religious systems. Sociality can inform how we understand artefacts themselves as well as how we can think through the set of socio-cultural relations from which textual meaning may emerge. No reader ever makes meaning alone, and no creator makes texts in isolation. This is especially true if scholars remember that all texts are connected to documents, whether those documents are written manuscripts, letterpress printed codexes, or ones and zeroes on magnetic discs. Inscription is unavoidable, and no inscription happens in a vacuum. By definition it leaves traces that lead us to a multiplicity of hands.

In 2001, and again in 2009, Jerome McGann has said that '(i)n the next fifty years the entirety of our inherited archive of cultural works will have to be re-edited within a network of digital storage, access, and dissemination'.⁵ While the time scale may be ambitious and seems more a provocation than a plan of action, the realities of cultural practice indicate that such a transition is well underway. Scholars have a valuable role to play in this collective remediation of our intellectual patrimony. To do so, however, individuals within academia must be present when decisions about funding, promotions, targets, and goals are discussed – and a key part of such discussions must be the nature of academic work in society at large. To remain and become more relevant, scholars must redefine scholarship as an expansive, welcoming process of collective cultural work. In short, I truly believe that for humanities work to survive as more than a set of disciplines increasingly under siege by legislators and dismissed by an underinvested public, humanists not only must accept that knowledge is and always has been social and that digital communications technology are changing the ways that knowledge work can be done, but embrace such a renegotiation of the scholarly landscape. Instead of standalone models of individual, isolated scholarship wherein which editors insist that control remain centralised in authoritative hands, humanists must grapple with the necessity of a scholarship that is public facing, that integrates diverse groups into creative knowledge – making activities, and that is social in the most positive sense. The Devonshire Manuscript Project and the Renaissance Knowledge Network are two attempts to enact this belief and are discussed at length elsewhere.⁶

It is worth remembering Williams' definition here, especially because of the way that 'social knowledge' as a theoretical construction fits so neatly into his narrative of objectifying and generalising relationships between actual people, much like the spate of 'social x-y-z' that formed throughout the 19th century. Humanists, collectively, often have placed internal practices of knowledge work into a black box closed to outside scrutiny, and even more so to outside involvement or investment. This has resulted in a false sense of separation, but one that has become so normalised that it no longer requires comment or explicit mention. In a larger dissertation project, I outline that one goal of the case studies, published content, and prototypes gathered together is to push back against some of the assumptions around critical work in the humanities, as well as advancing discussions about how

5 McGann 2009: 13.

6 See Powell *et al.* 2013; Powell *et al.* 2015 and Powell *et al.* 2014.

best to create and encourage social knowledge work in specific content areas.⁷ This follows logically from the idea of ‘the social’ as an analytical construct – as a way of interpreting texts and the world writ large – and the social as a way of practically and effectively re-mediating our ‘inherited archive of cultural works.’ These two meanings of the social – as analytical construct and as a way of discussing real-world interconnections facilitated by primarily digital means – are two sides of the same proverbial coin; namely ways of exploring social knowledge as practice within the contemporary humanities.

The dissertation mentioned above explores two of the major scholarly functions that digital humanists often find themselves undertaking: digital scholarly editing (remediating cultural content for publication, dissemination, and reuse online) and research infrastructure planning and execution (putting the systems in place that will facilitate future humanities research in self-reflexive ways). For both of these functions, social knowledge should be reconsidered so that, ironically, it would have made sense centuries ago: as a corollary to building fellowship and of fostering community. Approached in this way, social knowledge creation, whether it takes the form of editing, peer review, authorship, project building, or other activities within the ambit of digital humanists at work today, should not only be something to encourage and work towards, but also a way of remaking and remediating knowledge in practical and institutional terms. Scholarly communication cannot and does not exist in the absence of communities of practice that bring it in to being. And in this it is tied to material production, dissemination, and circulation of knowledge in concrete terms amongst those communities. Artefacts can productively reveal relations between embodied individuals, and subsequently be iteratively reconsidered to further foster such necessary ties.

7 Powell 2016.

References

- McGann, Jerome, 2009. 'Our Textual History.' *The Times Literary Supplement*. 20 November. Issue 5564. 13-16.
- Pierazzo, Elena, 2015. *Digital Scholarly Editing: Theories, Models and Methods*. London: Ashgate.
- Powell, Daniel. *Social Knowledge Creation and Emergent Digital Research Infrastructure for Early Modern Studies*. PhD., University of Victoria, 2016.
- Powell, Daniel, Constance Crompton, and Raymond G. Siemens.'Building the Social Scholarly Edition: Results and Findings from A Social Edition of the Devonshire Manuscript.' *Digital Humanities Abstracts*, 2013. <http://dh2013.unl.edu/abstracts/ab-300.html>.
- Powell, Daniel, Raymond G. Siemens, and William R. Bowen, with Matthew Heibert and Lindsey Seatter.'Transformation through Integration: The Renaissance Knowledge Network (ReKN) and a Next Wave of Scholarly Publication.' *Scholarly and Research Communication* 6, no. 2 (2015). <http://www.src-online.ca/index.php/src/article/view/199>.
- Powell, Daniel and Raymond G. Siemens, with the INKE Research Group.'Building Alternative Scholarly Publishing Capacity: The Renaissance Knowledge Network (ReKN) as Digital Production Hub.' *Scholarly and Research Communication* 5, no. 4 (2014). <http://src-online.ca/index.php/src/article/view/183>.
- Siemens, Raymond G., Karin Armstrong, Barbara Bond, Constance Crompton, Terra Dickson, Johanne Paquette, Jonathan Podracky, Ingrid Weber, Cara Leitch, Melanie Chernyk, Daniel Powell, Alyssa Anne McLeod, Alyssa Arbuckle, Jonathan Gibson, Chris Gaudet, Eric Haswell, Arianna Ciula, Daniel Starza-Smith, and James Cummings, with Martin Holmes, Greg Newton, Paul Remley, Erik Kwakkel, Aimie Shirkie, and Serina Patterson, with the Devonshire Manuscript Editorial Group. *A Social Edition of the Devonshire MS (BL Add 17, 492)*. http://en.wikibooks.org/wiki/The_Devonshire_Manuscript.
- Williams, Raymond. *Keywords: A vocabulary of culture and society*. Rev. ed. New York: Oxford University Press, 1983.

Beyond Open Access

(Re)use, impact and the ethos of openness in digital editing

Anna-Maria Sichani¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Introduction

'Open Access' is something of a buzzword in academia and in digital scholarly editing. My paper aims to critically engage with digital scholarly editing and Open Access through the lens of reuse and value creation. Such an approach distances itself from discussions of funding models needed to cover the costs of providing free access to expensive scholarship as well as from copyright and licensing issues. To be honest, it aims to get one step further: *beyond Open Access* as we used to think of. I first will attempt to describe how (what I call) an 'enlarged ethos of openness' succeeds in promoting creative reuse and redistribution of diverse digital content and data and how reuse is linked with and contributes to the value, impact creation and sustainability of digital data. I claim that such an approach of assessing value through reuse should be central in contemporary discussions of digital scholarly editing and Open Access, given that the majority of digital editions result from publicly funded projects, where the scholarly content and its value are need to be reconceived and redefined outside the sphere of monetary and material exchange, outside traditional economic/ market thinking.

In my main analysis, I will first try to map (though a qualitative and quantitative approach) the current practices employed in digital editions towards Open Access and reuse. Although the information environment in which digital scholarly content is created and delivered has changed phenomenally over the past fifteen years, allowing the sharing and reuse of digital data, and though the number of

¹ anna-maria.sichani@huygens.knaw.nl.

publicly accessible digital editions remains on the increase, I argue that limitations in adopting an Open Access agenda focused on reuse in digital scholarly editing still persist. I then will attempt to reveal the main dissuasive reasons for such a stance by questioning mainly the degree to which the patterns of reuse in scholarly editing have changed as we have moved from print to the digital. Finally, I will try to present and further discuss the different models of reuse in digital scholarly editing on both a theoretical and practical level, hopefully laying out a persuasive argument for the multiple benefits of such an endeavour.

The virtuous cycle of open digital content: re-use and value

As much as it gets: Open Access and the 'ethos of openness'

Open Access: what's in a word? Originating in the 'free' software movement and soon adopted in the early 2000s by academics and libraries calling for 'free immediate access to, and unrestricted reuse' (PLOS) of scholarly research, Open Access is grounded in a dual rupture as regards 'price and permission barriers' (Suber 2012). In the past decade, the culture and practices of Open Access have been expanded strategically and reinforced by the rhetoric of Open Definition, celebrating content/ data that is 'freely used, modified, and shared with anyone for any purpose' (Open Definition). Such an expansion and insistence on defining the 'openness' in the digital information environment marks the growth of open-related initiatives and communities of practice (e.g. Open Data, Open Source, Open GLAM) while also helping us uncover the major principles of 'openness': 1) Availability and Access, 2) Re-use and Redistribution, 3) Universal Participation.

Though the OA movement has made great strides in the last decade, it is important to celebrate all the small and big victories while also understanding that there is a complex matrix behind Open Access, meaning various degrees and combinations of the abovementioned principles. Speaking of barriers, there are two non-equivalent kinds of free online access: to adopt the terminology established for scholarly publications in journals, there is 1) 'gratis OA', which removes price barriers but not permission/copyright barriers, and 2) 'libre OA', which is free of both price barriers and unnecessary copyright and licensing restrictions, allowing reuse rights which exceed fair use/fair dealing enshrined in copyright law. There is room for variation here, as there is more than one kind of permission barrier to remove and more than one way to do it (type of re-use: copying, redistribution, derivative works, commercial re-use / attribution and right to remove attribution / indication of modification) and this is what Open Licenses (such as Creative Commons or Open Source Licenses) are designed to offer, i.e. a simple, standardized way of sharing and reusing works under a choice of conditions. Besides different types of barriers, there is also variety in what is actually open, speaking of output level and source-files and documentation level.

Open and useable digital content: value through reuse

Adopting such an ‘ethos of openness’ necessarily must reflect on the existing value systems of both academia and the market: in brief, it is mainly about disaggregating the concepts of value from cost and price as the conventional law of demand and supply and the scarcity principles teach us. Instead of making content valuable by making it scarce, as it is the rule in traditional economics and cost/market-based pricing, Open Access makes new knowledge/data valuable by making it widely available and open to reuse.

Such an understanding of value through (re)use, centres around the concept of non-rivalrous commodity exchange (Eve 2014; Suber 2012), introducing a remarkable dissonance with traditional evaluation paths for scholarly or cultural content. All in all, the idea of re-use could be used as an alternative metric with which to assess the value and impact of digital resources.

In light of the expanding mass of digital content created in the last decade, there also has been an ever-growing research interest in the area of use and impact assessment for open digital content, which is an incredibly difficult – if not impossible – task; many initiatives have developed and applied an array of both qualitative (stakeholder interviews, resource surveys, user feedback, focus groups, and questionnaires) and quantitative methods (e.g. webometrics, log file analysis, scientometric (or bibliometric) analysis, and content analysis) (TIDSR, Hughes *et al.* 2013) that capture ‘information about the whole cycle of usage and impact’ (Meyer *et al.* 2009: 6) in order to assess the usage and impact of digital resources (Warwick *et al.* 2006; Hughes 2012; Meyer *et al.* 2009; Tanner 2012).

Though this interest is mainly the result of pressure exerted by funding bodies (e.g. NEH, AHRC, Wellcome Trust, AHRC, EU/Horizon 2020, Mellon) and governments insisting on ‘the need to demonstrate the ‘impact’ of publicly funded resources and research, as a means of quantifying the value of the investment in their creation’ (Hughes 2012: 2), alongside guidelines of major funding bodies specifying that projects’ outputs (esp. software) should be ‘free in every sense of the term, including the use, copying, distribution, and modification’ (NEH 2013) and the opening up of many GLAM institutions datasets (such as DPLA, Rijksmuseum, Cambridge Digital Library; see also Terras 2015) for reuse and redistribution, it demonstrates nevertheless that stakeholders are now strategically moving from the creation of digital content towards a more sophisticated awareness of how notions of reuse, value, impact, open access and sustainability remain entwined.

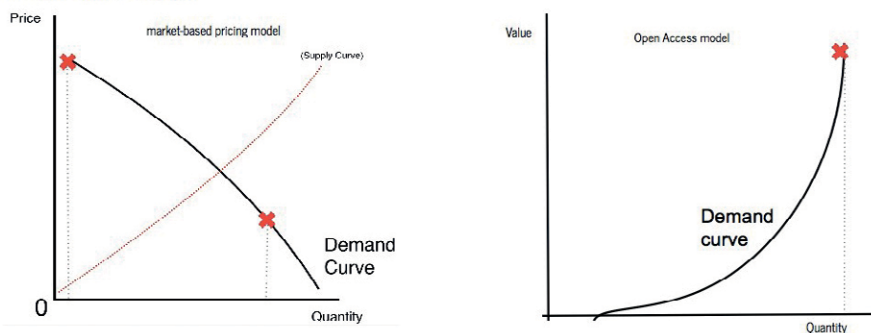


Figure 1: Law of demand / supply – Scarcity principle.

Digital Scholarly Editions made (not-so) open

Reading the numbers

How does this embracing of the virtuous cycle of open digital scholarly and cultural content and such an ethos of openness promoting ‘value through reuse’ relate to the field of digital scholarly editing?

In order to answer this question, I have collected data through qualitative analysis of digital editions projects and I also have extracted some quantitative data from Greta Franzini’s catalogue of Digital Editions² (total: 210 editions), to which I contributed.

Even though, historically, digital editions were conceived and developed with the aspiration of ‘*making ()’s vast work freely and conveniently accessible to scholars, students, and general readers*’, what digital editing endeavours actually achieve is more a critical curated digital reunification/gathering of materials scattered around the world in libraries and private collections. In other words, it succeeds in breaking down the textual scholarship barriers of time and space as regards the primary material. In addition, as a vast number of Digital Scholarly Editions initially are developed under grant cycles from funding agencies and governments, they have to conform from the outset to OA in the sense that they need to be free of charge. Despite persistent rhetoric regarding accessibility and availability, a relatively small percentage of those (only 12.7%) make their ‘beyond-the-output content’ openly available for reuse and redistribution.

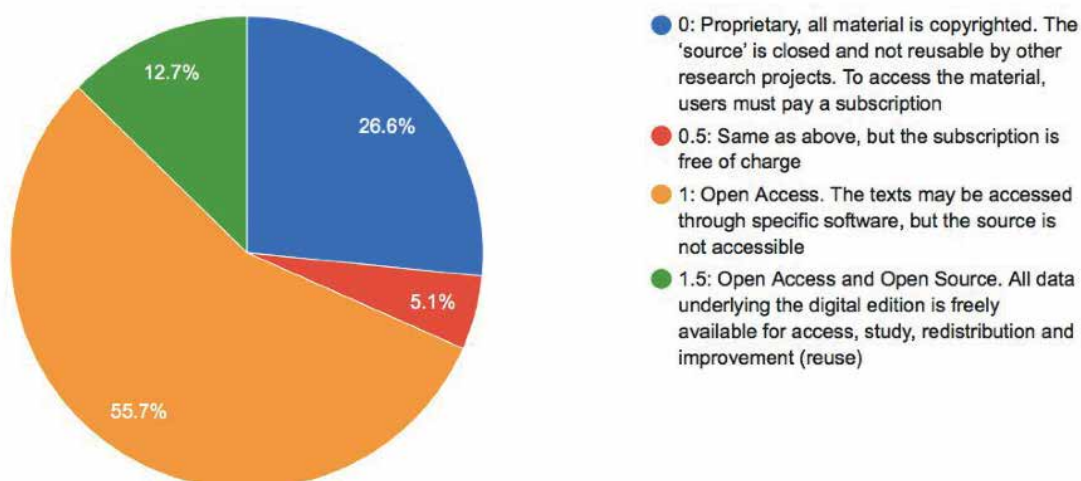


Figure 2: Catalogue of Digital Editions and Open Access.

² https://github.com/gfranzini/digEds_cat.

Revealing the barriers

Thus barriers still remain, both in the conception and in the practices of what it means for a digital edition to behave in an open access manner. The reasons for this may vary.

First of all, digital editions consist of cross-domain and heterogeneous primary material (e.g. text transcriptions, facsimiles, critical apparatuses, indices etc.) and thus issues of copyright may still persist or there may be different licensing frameworks for each of these materials (e.g. copyrighted facsimiles, pictures from the 20th and 21st centuries, orphan works etc.). Moreover, behind the scenes digital editions usually consist of far more than the output level: encoded XML source files, XML schemas, processing and transformation scripts, entity relation models, rich metadata, transcription and XML encodings, encoding conventions and decisions, project documentation (feasibility studies, proposal) and so on. All of them, are interpretative decisions or editorial choices and integral parts of the project's workflow.

Usually, as digital editions are developed within project-based frameworks, in which teams and contributors perform flexible roles, and allocated with fluid or overlapping tasks, claiming ownership or authorship of digital data is difficult to determine. In addition, a fear remains regarding losing control over your data (meaning that you are no longer able to update it) as well as a fear of malicious (re)use.

Aside from legal, economic, or operational reasons behind licensing complications, I strongly believe that the 'page paradigm' (Sahle 2008; Pierazzo 2015) remains a crucial hindrance in adopting an openness ethos that will encourage reuse; its inheritance is still so strong in our scholarly culture that we remain 'zoned to print' (Sutherland 2009: 20), thus tending to create and use digital editions 'as restrictedly repositories of data and generators of print editions' (Sutherland 2009: 2), as end-products or self-contained entities handed over to the end user 'to be seen and not touched' (Shillingsburg 2010; Dahlström 2011: 103). The 'look-but-don't-touch' problem, as Dahlström terms it, is rooted, on the one hand, in a strong scholarly attachment to completed, finished, publishable work, related to enduring issues of attribution of credit and professional reward. On the other hand, the problem is the partial result of a limited knowledge or evidence of how people might interact with and (re)use digital scholarly editions and their components.

Re-use' in the history of modern textual scholarship

I am firmly convinced that we deserve more than a read-only world – and in digital editing as well. Digital scholarly editions remain in an incunabular stage, primarily because we are too reluctant to change our habits when interacting with / using digital editions, which by definition possess different material qualities from printed ones. In other words, we continue to use them in much the same way as we use(d) a printed edition: for reading, consultation, observation, or in order to incorporate the new findings enabled by the computational technology into our next printed scholarly output (an article or a book). Such a pattern of reuse

differs little from the long established model of reuse in print culture: the footnote, the apparatuses and the related reference systems. While the footnotes ‘offer the empirical support for stories told and arguments presented’ (Crafton 1999: vii) and manifest a social model of knowledge creation, it also suggest a fertile but frozen interaction with the source, whose terrain of production and reproduction was defined by the print culture. As ‘electronic texts (and editions) are artefacts or mechanisms... amenable to material and historical forms of investigation, (thus) challeng(ing) textual critics to respond to the new medium in terms of its own materiality, architecture and functioning, as distinct from those of print’ (Sutherland, 2009, 22), I would suggest that the real advantages of digital editing will become apparent as we further understand and reuse digital scholarly editions in terms of their very own materiality and functioning, thus advancing us beyond ‘screen essentialism’ and exploring new patterns of re-use.

Towards an ethos of openness and reuse in digital editing

How we can reuse a digital edition?

There have been several theoretical proposals for how a re-use approach might be applied to digital editing. We can trace a few early but brave examples of such a practice.

In their proposal for *Open Source Critical Editions*, Bodard and Garcia (2009) support the distribution of the raw code, the documentation of the tools and applications that were used in reaching these conclusions, and the methodological statement behind the output, claiming that this is not only a commitment to operational transparency; it also amounts to a critical stance, in the sense that it ‘implies the adherence to reasonable methods and principles’. Getting ‘full access to the raw code is what that makes an experiment or a solution reproducible’ otherwise ‘it is a dead end; it cannot be built upon’.

In a similar vein, James Cummings (2009) claims that the model of ‘agile editions’ (easily transformed, re-purposable) based on the stand-off encoding proposal presupposes the opening up and distribution of underlying XML files alongside the HTML versions rendered. Moreover, by having first-hand examples of both successful and unsuccessful forms of community practice, we will be able to discover our joint errors and misunderstandings of the Guidelines and thus aid their improvement.

Peter Boot and Van Zudert’s proposal for ‘digital scholarly edition 2. 0’ (2011) suggests a decentralised, diverse and distributed architecture of openly available sources, services and functionalities (virtual research environment), initially enabled through cloud computing and cloud storage. This will reduce the amount of custom developed software and enhance the permanence of our digital outputs.

Models of reuse and/as aspects of value

Let me further extend the abovementioned abstract models by introducing how an open-ended model of deep access can be implemented in digital editing projects and mainly discuss the potential value of such a reuse ethos. Its foundations mainly

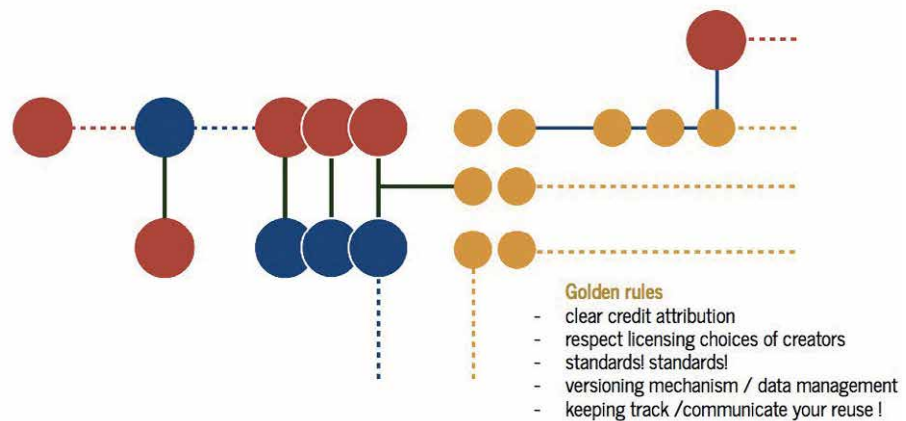


Figure 3: Open-ended deep access model.

rely on Dynamic systems development principles as found in the agile project delivery framework. In the open ended deep access model digital scholarly editions tend to be seen not as a static, closed-ended project or fixed product but more as an ongoing and open scholarly enterprise, as a continuum of iterative and incremental reuses of the various components of the project.

Deep access calls for opening up for reuse and redistribution source files (encoded XML, related schemas and ODDs, scripts for processing and visualisation, transformation scripts and query algorithms, entity relational models etc.). By going even deeper, we might also think about sharing parts of the project's documentation such as funding proposals (successful or not), feasibility studies, editorial conventions, even budget plans (as in most cases we are speaking of publicly funded projects) as supportive material for 'how to get a digital editing project off the ground'.

It is not the Open Access rhetoric per se but we need to find more elaborated ways to question, reveal and assess the value and the impact of such an ethos of openness in digital scholarly editing. By using some of the available quantitative methods to measure the impact of a digital edition (such as web analytics or log file analysis), we might gain important data about file requests, but it is extremely doubtful whether we are going to have any insight in the actual trajectory and value of its reuse. We need to engage in a more specialised and qualitative analysis of the patterns of reuse.

Aside from guaranteeing transparency, opening up and sharing source files also yields raw material for a new kind of research and scholarship that could be re-integrated and built upon it; an ideal test-bed for new computational/analytical approaches, either separate from or elaborated through a big data approach; an easy way to produce derivatives and varying outputs on demand (ePub, PDF) with potential for commercial exploitation; an opportunity for improvement and refinement; and finally, exemplary reference material for teaching purposes. An excellent example with such open 'fertile files' can be found at the *Jonathan Swift Archive*, a project that opens up for *download and reuse* its source files and scripts (TEI-XML, XSLT) so that researchers in the *Centrum voor Teksteditie en Bronnenstudie* (KANTL-CTB) have created a prototype with parallel readings of versions of Swift's works.

Reusing openly available source files also may contribute to reducing development costs, making future projects and their funding proposals more competitive. Though there, until now, has been a ‘black hole’ surrounding the economics of digital editing projects, I think we could easily claim a great potential for cost and time-avoidance: Latest statistics from Transcribe Bentham argue that not only could we save at least £400,000 if the remainder of the Bentham Papers were transcribed by volunteers, but that the *Collected Works of Jeremy Bentham* (which will run to approximately 70 volumes), could save up to 6 months of research staff time per volume (Causer *et al.* forthcoming). Additionally, the reuse of existing XML files contributed to a considerable reduction in the overall budget estimation, making more competitive and finally successful the proposal of the new AHRC funded Bentham project.

Finally, re-use and re-integration of existing data also may provide a distributed archiving solution (LOCKSS, Tapas), as well as a sustainability venue for continuous refinement and update, securing its long-term sustainability. By adopting such an approach familiar from the sustainable environmental management agenda: while reducing new demand for raw products, we succeed in avoiding depletion, to exploit the maximum of existing resources and thus to ensure the longevity of the planet, in a similar vein

Conclusion

To conclude, I think it is the right time to revisit Open Access in digital editing by focusing and further developing aspects of a more radical approach towards sharing, distributing and reintegrating digital content and outputs and the value of such an undertaking. I strongly support that it is not enough to claim for Open Access per se and struggling on the economical and distribution aspect but it would be more important to elaborate, develop and adopt practices and models that will help us to have more Open, Usable, Used and Useful digital scholarly editing future.

References

- A catalogue of Digital Scholarly Editions* (v 3. 0, since 2008ff), compiled by Patrick Sahle, (last change 30 December 2016). Accessed March 5, 2017. <http://www.digitale-edition.de/vlet-about.html>.
- Bodard, Gabriel, Juan Garcés, Marilyn Deegan, and Kathryn Sutherland. 2009. 'Open Source Critical Editions: A Rationale.' In *Text Editing, Print, and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland, 83-98. New York: Routledge (Digital Research in the Arts and Humanities).
- Boot, Peter, and Joris van Zundert. 2011. 'The Digital Edition 2.0 and The Digital Library: Services, Not Resources.' In *Digitale Edition Und Forschungsbibliothek, Beiträge Der Fachtagung Im Philosophicum Der Universität Mainz Am 13. Und 14. Januar 2011*, edited by Christiane Fritze, Franz Fischer, Patrick Sahle and Malte Rehbein, 141-152. Wiesbaden: Harassowitz (Digitale Edition und Forschungsbibliothek).
- Causser, Tim, Kris Grint, Anna-Maria Sichani and Melissa Terras. Forthcoming. 'Making such bargain': Transcribe Bentham and the quality of cost- effectiveness of crowdsourced transcription'. *Digital Scholarship in the Humanities*.
- Crafton, Antony. 1999. *The Footnote. A Curious History*. Cambridge (Mass.): Harvard University Press.
- Cummings, James. 2009. 'Converting Saint Paul: A New TEI P5 Edition of The Conversion of Saint Paul Using Stand-off Methodology.' *Literary and Linguistic Computing* 24 (3), 307-317.
- Dahlström, Mats. 2011. 'Editing Libraries', *Bibliothek und Wissenschaft* 44, 91-106.
- Eve, Martin Paul. 2014. *Open Access and the Humanities, Contexts, Controversies and the Future*. Cambridge: Cambridge University Press.
- Hughes, Lorna M. 2012. 'Introduction: the value, use and impact of digital collections'. In *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, edited by Lorna Hughes. London: Facet Publishing, 1-12.
- Hughes, Lorna M., Paul S. Ell, Gareth A. G. Knight and Milena Dobрева. 2013. 'Assessing and measuring impact of a digital collection in the humanities: An analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project'. *Digital Scholarship in the Humanities* 30 (2), 183-198.
- Meyer, Eric T., Kathryn Eccles, Michael Thelwall and Christine Madsen. 2009. *Final Report to JISC on the Usage and Impact Study of JISC-funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Resources* (TIDSR). Accessed March 3, 2017. http://microsites.oii.ox.ac.uk/tidsr/system/files/TIDSR_FinalReport_20July2009.pdf.
- NEH (*National Endowment for the Humanities*). *Office of Digital Humanities, Digital Humanities Implementation Grants*. Accessed March 3, 2017. <http://www.neh.gov/grants/odh/digital-humanities-implementation-grants>.
- Open Definition*. Accessed March 4, 2017. <http://opendefinition.org/>.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing. Theories, Models and Methods*. London: Ashgate.

- PLOS (Public Library of Science)*. Accessed March 4, 2017. <https://www.plos.org/open-access/>.
- Shillingsburg, Peter. 2010. 'How literary works exist: implied, represented, and interpreted'. In *Text and genre in reconstruction: effects of digitalization on ideas, behaviours, products and institutions*, edited by Willard McCarty. Cambridge: Open Book Publishers, <http://books.openedition.org/obp/658>.
- Suber, Peter. 2012. *Open Access*. Cambridge (Mass.): MIT Press.
- Sutherland, Kathryn. 2009. 'Being critical: paper-based editing and the digital environment'. In *Text Editing, Print and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland, 13-26. London: Ashgate.
- Tanner, Simon. 2012. *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*. King's College London. Accessed March 5, 2017. www.kdcs.kcl.ac.uk/innovation/impact.html.
- Terras, Melissa. 2015. 'Opening Access to collections: the making and using of open digitised cultural content'. *Online Information Review* 39 (5), 733-752.
- Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. Accessed March 4, 2017. <http://microsites.oii.ox.ac.uk/tidsr/>.
- Warwick, Claire *et al.* 2006. *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities. Final Report to the Arts and Humanities Research Council. School of Library, Archive and Information Studies*. London: University College London. Accessed March 5, 2017. <http://www.ucl.ac.uk/infostudies/claire-warwick/publications/LAIRAHreport.pdf>.

The business logic of digital scholarly editing and the economics of scholarly publishing

Anna-Maria Sichani¹

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

Over the past two decades, scholarly discussions have been populated by a narrative of immanent crisis in the future of scholarly publishing, and especially in the Humanities. This crisis narrative affects the whole spectrum of scholarly communication (academics, libraries and publishing entities) and usually is associated with a complex mix of financial recessions, continuous market shrinkages, price increases, and the advent of electronic forms of knowledge production and communication. From infrastructure and technological issues, agents, assessment and evaluation criteria, to communicative structures, business models and impact frameworks, the academic community invests its efforts in the cultivation of an ever-growing arena of discussion and experimentation in a constant attempt to remodel roles and practices alongside various scholarly outputs (monograph, journals, dissertation) in the digital publishing ecosystem in terms of economics and long-term sustainability (MLA 2002; Alonso *et al.* 2003; Fitzpatrick 2011).

Meanwhile, textual scholars have been pioneering explorers and challengers of the changes introduced by digital technologies in textual production, representation and transmission, resulting in an ever-growing body of research on concepts, encoding standards, digital tools and methods alongside a variety of exemplary digital scholarly editions. Though digital editing typically has constituted a very creative and productive branch of digital scholarship, scholarly editing surprisingly has failed to find a space of participation in the abovementioned 'crisis discussion'. This comes as no surprise: digital editions produced so far primarily are developed

¹ anna-maria.sichani@huygens.knaw.nl.

as ‘exploratory loci’ – in other words, as ‘experiments, a way to test what is possible within the new medium and to establish new ways for scholarship’ (Pierazzo 2015, 204) and secondly as ‘products’ within the scholarly publishing and communication circuit, with distinct commodification aspects or features (e.g. cost, exchange- and use-value etc.). To put it differently: digital scholarly editing is now struggling to move out of a relatively sheltered environment operating at the pace of and through the grant-based funding of the academic enterprise, into one that operates at the speed of digital scholarly publishing and communication and in accordance with the new rules of web economy. While there are some exemplary cases of business mindsets informing digital scholarly editing projects, such an approach remains somehow fragmented within digital editing cycles. In what follows, I aim to propose some basic pillars of such a business logic.

Business model approach and sustainability

Given that the majority of digital editions so far have been linked organically with a research project logic, they programmatically have failed to address issues of business modeling, operating architecture and sustainability planning. However, the inherent technological qualities of various media types of digital data – components of digital editions – demand perpetual curation (both in technological and in scholarly terms) and thus related costs – a parameter rarely addressed within the project-based framework of digital editions. This partially could explain why digital editions created in the previous decades, considered monumental in terms of scholarly integrity and quality, are now obsolete in terms of technical infrastructure, functionality and/or interface. They now require a solid and costly maintenance/upgrade plan in order to re-state their place in the scholarly communication ecosystem.

A comprehensive business-driven approach to digital editing can aid our enterprise with crucial operational aspects such as fundraising avenues, project management and financial planning while informing our practices with a vast repertoire of cost-reduction strategies (e.g. out/crowdsourcing), revenue-generation streams (e.g. subscription, freemium, advertising etc.) and distribution channels, currently implemented and thriving in digital publishing and web commerce. By approaching digital editing projects’ sustainability through such a critical business mindset, I claim that we can challenge the narrative that persistently posits publicly funded and Open Access by principle digital resources in the Humanities as incompatible with the sphere of monetary exchange.

Value and impact creation

While the existence of different audiences within digital editing cycles is discussed frequently, mainly in relation to encoding strategies and resulting functionalities, such discussions are not always associated with a sophisticated demand-driven approach in terms of customizer-to-user needs formats and features. By implementing the principles of market research and segmentation in digital editing projects, we can thoughtfully design and ensure the access to and the value

– financial or otherwise – of our digital editions for targeted categories of users in the long-term.

Multiple formats and modalities (e.g. tablet version, apps, print version), advanced functionalities (e.g. print-on-demand, customized visualisation etc.), specialised tools and working environments – all the abovementioned strategies genuinely reframe Vanhoutte’s idea of the minimal and maximal edition (Vanhoutte 2012) through a market-oriented perspective: they offer a wider dissemination of our scholarship and act as an important revenue generator, while strategically enhancing the impact and the creative (re)use of the digital scholarly editions’ content. Concepts such as the reusability and the reproducibility of digital editions (Dahlström 2009) do not only describe our academic endeavour of knowledge transfer and scholarly exchange; they are also substantially linked to the economics of the digital edition, introducing cost avoidance solutions and additional revenue streams.

Furthermore, by adopting a hybrid concept of digital editions as an endlessly flexible scholarly resource and a customizable bundle of services that often genuinely exceed the scholarly realm, we succeed in attracting, retaining and expanding an audience while nourishing the interest and the value proposition of the resource.

Partnerships that count

Digital editing historically has been a collaborative enterprise between scholarly editors, libraries and cultural heritage institutions, university presses and technology specialists. Such a partnership ethos lies at the heart of digital scholarship and is vital for the successful accomplishment of highly demanding and expensive digital editing projects.

However, besides their cost-efficient purpose, well-planned partnerships also ensure, through a decentralization/division of labour, that each partner pointedly will contribute to his/her specialty in the long-term. Such a strategy results in high quality specialised services that will guarantee the quality as well as the afterlife of the digital edition in terms of maintenance, distribution and communication.

While the implementation around digital scholarly editing and publishing alongside their financial politics are still (and probably will remain) constantly in flux, I stand up for the experimentation with and adoption of more entrepreneurial and business-like mindsets in all aspects of such an enterprise. As digital editing is fighting to slowly transform itself from an experimental undertaking to an accepted and legitimate form of scholarship, it is also necessary to explore and further develop flexible and sustainable business models and procedures that will ensure the scholarly quality while broadening the communicative, educational and cultural role of the scholarly edition in the new ecology of digital scholarly publishing and communication.

References

- Alonso, Carlos J., Cathy N. Davidson, John M. Unsworth, Lynne Withey, 2003. Crises and Opportunities: The Futures of Scholarly Publishing, *American Council of Learned Societies ACLS Occasional Paper*, No. 57 ACLS Annual Meeting on May 10, 2003, https://www.acls.org/uploadedFiles/Publications/OP/57_Crises_and_Opportunities.pdf.
- Dahlström, Mats, 2009. 'The Compleat Edition', In *Text Editing, Print, and the Digital World* (Aldershot: Ashgate, 2009), edited by M. Deegan and K. Sutherland, 27-44 (originally published as 'How Reproductive is a Scholarly Edition?' In *Literary and Linguistic Computing*, Vol. 19, No. 1, 2004, 17-33).
- Fitzpatrick, Kathleen. 2011. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. NYU Press.
- MLA Ad hoc Committee on the Future of Scholarly Publishing, 2002. 'The Future of Scholarly Publishing,' *Profession*, (December 2002), <https://apps.mla.org/pdf/schlrlypbshng.pdf>
- Pierazzo, Elena, 2015. *Digital Scholarly Editing. Theories, Models and Methods*. Ashgate.
- Vanhoutte, Edward, 2012. 'Being Practical. Electronic editions of Flemish literary texts in an international perspective', *International Workshop on Electronic Editing (9-11 February 2012)*, School of Cultural Texts and Records, Jadavpur University, Kolkata, India, available at <http://edwardvanhoutte.blogspot.be/2012/02/being-practical-electronic-editions-of.html>.

The social edition in the context of open social scholarship

The case of the Devonshire Manuscript (BL Add Ms 17, 492)

Ray Siemens^{1, 2}

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Excerpted from: '*Building A Social Edition of the Devonshire Manuscript*.' Constance Crompton (University of British Columbia, Okanagan), Daniel Powell (University of Victoria), Alyssa Arbuckle (University of Victoria), and Ray Siemens (University of Victoria), with Maggie Shirley (University of British Columbia, Okanagan) and the Devonshire Manuscript Editorial Group. *Renaissance and Reformation* 30.4 (2015): 131-156.

A Social Edition of the Devonshire Manuscript is an unconventional text in that it attempts to bring traditional scholarly editing practices and standards into conversation with comparatively recent developments in online social media environments. In doing so, the edition aims to embody contemporary editorial theories recognizing the inherently social form and formation of texts alongside the social practices of writing, revision, and editing that shaped the original production of the Devonshire Manuscript (BL MS Add. 17, 492). Dating from the 1530s to the 1540s, the Devonshire Manuscript is a collaborative verse miscellany authored and compiled by a number of 16th century contributors.³ We believe that,

1 siemens@uvic.ca.

2 For Constance Crompton, Daniel Powell, Alyssa Arbuckle, Maggie Shirley.

3 Following Peter Beal's definition of a verse miscellany as 'a manuscript, a compilation of predominantly verse texts, or extracts from verse texts, by different authors and usually gleaned from different sources' in *A Dictionary of English Manuscript Terminology, 1450-2000* (London: Oxford University Press, 2008), 429. Beal lists the Devonshire Manuscript as a pertinent example of a verse miscellany in Beal, *Dictionary*, 430.

as an inherently collaborative text, the manuscript calls for an innovative approach to scholarly editing. In this article, we detail the content, context, process, and implications of *A Social Edition of the Devonshire Manuscript*.

A Social Edition of the Devonshire Manuscript is an innovative project, but one with deep roots in ongoing Canadian scholarship on Renaissance literature, scholarly editions, and digital humanities prototyping. Much of this transdisciplinary work has taken place under the aegis of two groups: the Electronic Textual Cultures Lab (ETCL) at the University of Victoria and the Canada-wide Implementing New Knowledge Environments (INKE), both directed by Ray Siemens.⁴ The ETCL engages deeply with the study of textual communication in all its historical, present, and future forms. Alongside this research mandate, the ETCL serves as a Vancouver Island-based hub for regional, national, and global digital humanities work and training; the highly successful Digital Humanities Summer Institute (DHSI) held annually at the University of Victoria is perhaps the flagship initiative of the wider digital humanities community.⁵ With graduate student researchers, postdoctoral fellows, affiliated faculty, visiting speakers, and regular community events, the ETCL is a vibrant research collective engaged in the wider examinations of the types of intellectual issues prompted by *A Social Edition of the Devonshire Manuscript*. As a research and prototyping initiative, INKE describes itself 'as an interdisciplinary initiative spawned in the methodological commons of the digital humanities that seeks to understand the future of reading through reading's past and to explore the future of the book from the perspective of its history.'⁶ Divided into three research areas – textual studies, modelling and prototyping, and interface design – INKE members interrogate the nature of textuality in the digital age. To date, the various INKE groups have produced a number of publications, sponsored several conferences, and built numerous digital tools and prototypes for scholarly use.⁷

As these brief synopses might indicate, the intertwining research communities present around the ETCL and INKE provide context for *A Social Edition of the Devonshire Manuscript*. The prototyping and consideration of what a digital, *social*, scholarly edition might look like is an expression of longstanding and well-funded Canadian research into the nature of the book in a digital age. It also attempts to put into practice Ray Siemens' argument that social media environments may enable new editing practices, itself an argument formulated in an article emerging from the collaborative research environment of the ETCL (Siemens, 445-461).

By publishing on Wikibooks (a partner site to Wikipedia focused on book-length projects) we emphasize the collective, social ethos of the original document itself. Throughout this process we have attempted to model the *social scholarly edition* and address the questions a social edition, and social editing, raise: How do we effectively integrate multiple communities with varying cultures and editorial

4 Website of the ETCL: <http://etcl.uvic.ca/>. Accessed 5 November 2014. Website for INKE: <http://inke.ca/>. Accessed 5 November 2014.

5 Website for DHSI: <http://dhsi.org/>. Accessed 5 November 2014.

6 'About,' INKE, <http://inke.ca/projects/about/>. Accessed 5 November 2014.

7 For publications, see: <http://inke.ca/projects/publications/>; for featured tools and prototypes, see: <http://inke.ca/projects/tools-and-prototypes/>; for conferences see: <http://inke.ca/skill/research-activities-engagement/>. Accessed 5 November 2014.

standards while pushing the boundaries of editorial authority? How do we employ various social media platforms with different degrees of openness to ensure a safe space for numerous individuals and opinions? And how do we shift the power from a single editor who shapes the reading of any given text to a group of readers whose interactions and interpretations form a new method of making meaning out of primary source material? To attend to these questions, this article begins with a description and consideration of the document itself – BL MS Add.17,492. Next, we recount the processes involved with building a digital social edition of this idiosyncratic text. To conclude, we interrogate the affordances and drawbacks of digital scholarly editing in collaborative, Web 2.0 contexts.

In order to build a scholarly edition on the principles of open access and editorial transparency (in both production and dissemination), we have integrated early modern content and scholarly editing practices with web-based environments maintained by established social and social-editorial communities – most notably on Wikibooks, a cross-section of intellectual research activity and the social media practices that define Web 2.0.⁸ Early on, Web 2.0 was described as internet technologies that allow users to be active authors rather than simply readers or consumers of web content. (*cf.* DiNucci 1999: 221-222) Now, the term is associated most frequently with social media platforms, wikis, and blog applications. As Tim Berners-Lee remarks, the internet originally was developed for workers to collaborate and access source documents; with wiki and Web 2.0 technology, it is now returning to its roots. (*cf.* Mahony 2011) The successful group of Wikimedia projects (Wikipedia, Wikibooks, Wikiquote, etc.) emphasizes the importance of multi-authored and multi-edited endeavours. In doing so, Wikibooks instantiates earlier theoretical arguments that texts are created by a community of individuals; as Marotti argues, ‘production, reproduction, and reception are all socially mediated’ (Marotti, 212). To put this into practice, we extended our editorial conversations into multiple pre-existing Web 2.0 and social media platforms, including Twitter, blogs, Wikibook discussion pages, dedicated Renaissance and early modern online community spaces, and Skype-enabled interviews with our advisory group. In creating *A Social Edition of the Devonshire Manuscript* we bring both Web 2.0 and current editorial theories of social textuality and community editing into closer focus. What is the outcome of scholarly editing if, like the originary Devonshire Manuscript contributors, we understand and enact the edition-building process as inherently collaborative? In what follows we offer a brief overview of the methods, process, and thinking that led to the Wikibook instantiation of *A Social Edition of the Devonshire Manuscript*.

Perhaps more than any other editorial choice, the iterative publication of *A Social Edition of the Devonshire Manuscript* departed most clearly from traditional scholarly editing practices. In effect we have published (or are in the perpetual process of publishing) two versions of the edition: a PDF version, distributed to the project’s advisory board; and a version housed on the publicly-editable

8 Wikibooks is a Wikimedia project that continues the aim of Wikipedia; namely, to encourage, develop, and disseminate knowledge in the public sphere. Wikibooks differs from other Wikimedia projects in that it is designed primarily for facilitating collaborative open-content textbook building.

Wikibooks. We currently also are working with multiple publishing partners to produce versions of the edition in other mediums: an SQL-backed edition on Iter: Gateway to the Middle Ages and Renaissance; an e-reader edition designed for tablets; and a print edition, published by the *Medieval and Renaissance Text Society*. Taken together, these multiple platforms can meet the needs of a broad and varied readership while, for the most part, growing organically out of a central set of texts and practices. These versions were planned to productively inform and influence each other's development, with cross-pollination of editorial input across platforms.

The Wikibook edition pushes the limits of what a print edition realistically can achieve – including in sheer size. Even if the manuscript facsimile pages and the XML files were excluded, *A Social Edition of the Devonshire Manuscript* would run to over five hundred standard print pages.⁹ In addition to a general and textual introduction, the online edition includes extensive hand sample tables that open our palaeographic attribution process to public scrutiny; witnesses that reflect the poem's textual legacy; biographies and genealogical diagrams that clarify the relationship between the manuscript's 16th century compilers; and an extensive bibliography of quoted and related sources. Courtesy of Adam Matthew Digital, we also have included the facsimile image of each page of the manuscript alongside transcribed content and explanatory notes. Going further, the discussion sections on each wiki page allow conversation on each item. The Wikibook edition extends the social context of the Devonshire Manuscript by providing a space for ongoing discussion and collaboration.

Editorial processes of *A Social Edition of the Devonshire Manuscript* began long before selecting Wikibooks as a publication platform. In 2001, work on a digital edition of the manuscript began with a more recognizably traditional scholarly activity: primary source transcription. The base transcription is based on examination of both the original document at the British Library and a microfilm of the Devonshire Manuscript, also provided by the British Library. Members of the Devonshire Manuscript Editorial Group (or DMSEG, a team made up of scholars, postdoctoral fellows, graduate researchers, and programmers,¹⁰ working

9 The DMSEG did, in fact, export the Wikibook edition to print format in summer 2013; the two-volume, hardback edition is approximately 1000 pages.

10 Ray Siemens, Karin Armstrong, Barbara Bond, Constance Crompton, Terra Dickson, Johanne Paquette, Jonathan Podracky, Ingrid Weber, Cara Leitch, Melanie Chernyk, Brett D. Hirsch, Daniel Powell, Alyssa Anne McLeod, Alyssa Arbuckle, Jonathan Gibson, Chris Gaudet, Eric Haswell, Arianna Ciula, Daniel Starza-Smith, and James Cummings, with Martin Holmes, Greg Newton, Paul Remley, Erik Kwakkel, Aimie Shirkie, and the INKE research group.

with two publishers¹¹, an editorial board¹², and self-selected members of the public) prepared and transcribed (in a blind process) two independent transcriptions from the microfilm. The transcribers collated the two paper copies manually, and the resultant rough text was resolved as far as possible using expanded paper prints and enlarged images. Their transcriptions were largely in accord with one another. Remaining areas of uncertainty were resolved with manual reference to the original document itself. This final, collated transcription forms the textual basis for *A Social Edition of the Devonshire Manuscript*, the basis of all editorial activity.

Following this process, the team then encoded the text in XML according to Text Encoding Initiative (TEI) guidelines.¹³ While encoding, the team upheld principles of consistency and accountability. Even if the team discovered a choice to be less than optimal, they continued in that pattern until the text was complete. Rather than employ varying practices, consistently encoding the entire manuscript in XML allowed for global changes that could be, and indeed were, made after the conclusion of the initial encoding.¹⁴ Furthermore, while encoding the team maintained regular documentation to ensure that neither the original encoder nor any subsequent encoder would lack a basis from which to proceed. Another practice employed was to encode the manuscript by building layers of TEI in phases. The manuscript was encoded completely at a conservative level before commencing the second phase. The second layer of encoding, complete with annotations and regularizations, deepened, clarified, and augmented the first. This tiered process also allowed for the encoding of doodles, anagrams, and other non-textual materials found within the manuscript.

Although the project began in 2001, the social edition on Wikibooks started with the formation of an advisory group in 2010. Throughout the production of *A Social Edition of the Devonshire Manuscript*, we consulted and conducted qualitative interviews with members of this advisory group to gather their perspectives on the content of the evolving edition. Forming an advisory group provided a unique opportunity to invite potential users and reviewers to shape the process and products associated with the social edition. As the final step before moving the text to Wikibooks, the members of the DMSEG working in the ETCL prepared a static digital edition of the manuscript. This fixed edition served as a base text against which our international advisory group of early modern and Renaissance scholars could compare the Wikibooks edition as it evolved.

11 Iter, a not-for-profit consortium dedicated to the development and distribution of scholarly Middle Age and Renaissance online resources in partnership with Medieval and Renaissance Texts and Studies and Adam Matthew Digital, a digital academic publisher.

12 Robert E. Bjork (Director, Arizona Center for Medieval and Renaissance Studies; Arizona State University), William R. Bowen (Chair) (Director, Iter; University of Toronto Scarborough), Michael Ulliyot (University of Calgary), Diane Jakacki (Georgia Institute of Technology), Jessica Murphy (University of Texas at Dallas), Jason Boyd (Ryerson University), Elizabeth Heale (University of Reading), Steven W. May (Georgetown College), Arthur F. Marotti (Wayne State University), Jennifer Summit (Stanford University), Jonathan Gibson (Queen Mary, University of London), John Lavignino (King's College London), and Katherine Rowe (Bryn Mawr College).

13 TEI provides a standard for encoding electronic texts. By encoding a text in XML under TEI guidelines, one renders the text substantially more searchable, categorizable, and preservable.

14 Please note that these global changes were not questions of textual transcription, but of encoding patterns and standards.

Before deciding on Wikibooks as a platform, the team had considered hosting the edition on a stand-alone site. In response to public interest in the project, coupled with the team's investment in emerging public knowledge communities, we instead developed a two-pronged strategy: as a control we produced a static PDF version of the edition, and as a variable we moved the same content onto a Wikimedia platform. Most famous for Wikipedia, Wikimedia is a small non-profit foundation responsible for management, fundraising, and technological development of Wikipedia, Wikibooks, Wikisource, Wiktionary, and a number of other projects. Volunteer editors contribute and moderate the content of all projects with self-developed norms and systems of oversight. We considered Wikisource, Wikibooks, and Wikipedia as platforms, eventually deciding to mount our edition in Wikibooks. Acknowledging the dedicated community already engaged in Wikimedia, we sought to discover Wikibooks' affordances for the scholar. Even though Wikipedia has far more editors, Wikibooks purposefully is structured to support the book-like form. And although Wikisource may appear a more appropriate environment for an edition, publishing *A Social Edition of the Devonshire Manuscript* on Wikisource would have disallowed the inclusion of any and all scholarly material outside the transcription itself – including paleographic expansions, appendices, notes, and bibliographies. With a book-like resource as our end goal, we produced a scholarly and peer reviewed edition in Wikibooks that also enables citizen scholars to access, contribute, and annotate material. Crucially, Wikibooks also archives each change in any content, allowing us to track reversions and revisions to the text.

In order to keep the editorial and encoding process transparent, the Wikibook edition includes links to the baseline XML-encoded transcription. In addition to being able to use the XML for their own projects, readers conversant with XML can see the encoder's TEI-based editorial choices. Anyone can download this XML and continue working with the XML in any way they see fit, allowing the project to potentially evolve in unanticipated ways.¹⁵ With the firm foundation of documented encoding, all those working with the document can refer to, build on, or adapt the project's foundation. Readers can compare our transcriptions to the facsimiles included on each page of the Wikibooks edition and are free to contest (and even alter) our regularizations or corrections.

In November 2011, ETCL-based members of the DMSEG began converting the TEI-encoded text into Wikimarkup, the unique language designed for wiki publication. The team then moved the text, appendices, glosses, commentary, and textual notes into Wikibooks. Wikibooks, like Wikimedia and institutional scholarship at large, has its own self-governing editorial culture, and *A Social Edition of the Devonshire Manuscript* received attention from Wikibooks' existing editorial community. Since then, the ETCL team has amplified the base text with additional images of the manuscript, witness transcriptions, an extensive bibliography, and the XML files containing the encoded transcription of the manuscript. Consequently, the Wikibook became a hybridized edition-research environment for both early modern scholars and Tudor enthusiasts. Various

¹⁵ These can be downloaded from the Wikibooks edition site.

authors have written on these phenomena, and on the value of employing wikis as collaborative research or authoring platforms; best practice standards and protocols have developed as an increasing number of researchers (both academic and not) become versed in Wikipedia methods. We consciously have developed *A Social Edition of the Devonshire Manuscript*, a scholarly Wikibook edition, with these practitioners, priorities, and standards in mind.¹⁶

The Wikibook platform gives us the opportunity to recognize and assign credit for important editorial work that extends beyond the creation of original base text. Activities like discussion and feedback are central to scholarly revision and authorship, but can be difficult to monitor and quantify in a large project. A print edition often only acknowledges these forms of labour with a line or two on the acknowledgments page. Originally, we considered the discussion pages ideal for this type of scholarly discussion and editorial record keeping. Like any private community, however, Wikibooks bears its own social conventions. Through conversation with an established Wikibooks editor we realized that the Wikibooks discussion pages are used more often for personal commentary and disputes than editorial suggestions. Reminiscent of Douglas's note in the margin of 'Suffryng in sorrow in hope to attyn' (fol. 6v – 7r) to 'fforget thys, ' and Shelton's contradiction 'yt ys wor(t)hy, ' these pages are predominantly venues for editors to offer one another personal support (or criticism) rather than to analytically discuss content in a way scholars might find useful in a research context. Although the technology readily supports our original intention, the cultural practices of the Wiki community required us to alter our expectations. Despite this, all edits to all pages of the project are recorded on each page's 'View History' tab.

Thus, rather than relying on the discussion pages for editorial debate and decisions, we made the most substantive changes in Wikibooks based on Skype and Iter interactions with our advisory group. Although our hope had been to have the advisors edit directly in Wikibooks, many found the technological threshold for contributing too high, and it became more practical to have the ETCL team make the proposed changes to WikiCode. We responded to the advisors' recommendations in near-real time, adding (among other suggestions) navigation menus and facsimile page images. This is, again, a cultural issue rather than a technical one: the social edition has always been, and remains, open for anyone to edit at any time. Short of locking a page by an administrator (an action often undertaken only for repeated vandalism or during edit wars), there is no mechanism for denying anyone the ability to edit. As we found, many avenues for editorial conversation are necessary in order to foster the sense of a community that, as one of our advisors noted, is 'virtually there, as if everyone is crowded around a page, putting their two cents in on matters great and small'. Even when those giving editorial direction do not directly make changes to the edition, the use

16 Bo Leuf and Ward Cunningham, authors of the first book on wikis, recognize that a wiki must fit the culture of the user community for it to be successful (Leuf and Cunningham 2001). Emma Tonkin advises that a collaborative authoring wiki should include the following: a page locking system to deter simultaneous editing, a versioning system to track changes, and the ability to lock editing on a page in the case of an edit war, as well as an efficient search function, and navigation, categorization, and file management abilities (Tonkin 2005).

of multiple social media platforms like blogs and Twitter can productively facilitate social editing discussions. Focusing solely on one single communications platform potentially could impede the success of an evolving edition.

As we discovered, every social media platform attracts and enables specific types of interaction. Using social media allows us to integrate a new step into the editorial process – a step that fills the gap between initial planning stages and concluding peer review reports. Producing an edition ‘live’ in consultation with various groups across multiple media engenders an edition that quickly and productively meets the needs of its readers. Employing and participating in various platforms alerted us to different priorities across platforms, as well as forcing us to think through how we might create a polyvocal experience for safe, productive, and equitable interactions.

In addition to producing an edition that allows for multiple editorial perspectives, the DMSEG gathered responses to the social edition-building methodology. In the interest of refining the process and expounding on its utility for collaborative editors in the Web 2.0 environment, the ETCL team used a combination of methods to gather data on the social edition-building process. We invited feedback via Twitter, guest blog posts, and Iter’s social media space. We also encouraged direct intervention in the Wikibooks edition of *A Social Edition of the Devonshire Manuscript*. Furthermore, we consulted with members of our advisory group on issues of credit, peer review, and collaborative decision-making. Rather than soliciting anonymous reader reports from our advisors, we brought them into conversation with one another over the fixed edition and the evolving Wikibooks edition. We facilitated this conversation in a social media space housed by Iter, a federated site housed at the University of Toronto that serves a broad community of early modern and Renaissance associations and scholars.¹⁷ In many cases, their suggestions already have been incorporated into the Wikibooks publication; those that have not will be integrated into a final, socially-produced edition of the Devonshire Manuscript for print and e-publication with Iter and Medieval and Renaissance Texts and Studies (MRTS).¹⁸

Considered as a whole, *A Social Edition of the Devonshire Manuscript* suggests that social media technologies can be harnessed for productive interaction and discussion by those scholars invested in a content area or project, but that they require comprehensive oversight by dedicated staff to develop and maintain participation in knowledge construction and dissemination. Regardless, social scholarly editions represent a step towards diversifying and democratizing knowledge, and the Wikimedia suite of platforms is an established environment

17 See Iter: Gateway to the Middle Ages and Renaissance at <http://itergateway.org/>. Accessed 5 November 2014.

18 These various avenues of participation met with different levels of success, the overview of which is outside the scope of this article. Our team has presented on this aspect of the project at Digital Humanities 2013 (see <http://dh2013.unl.edu/abstracts/ab-300.html>), and a forthcoming article focuses more intently on stakeholder communities and their responses to the project. See Constance Crompton, Raymond Siemens, Alyssa Arbuckle, the Devonshire Manuscript Editorial Group, and INKE ‘Enlisting ‘Vertues Noble & Exceleent’ Across Scholarly Cultures: Digital Collaboration and the Social Edition.’ *Digital Humanities Quarterly* (accepted).

for this sort of work. Todd Presner reiterates this concept by considering Wikipedia as a model for the future of humanities research, deeming Wikipedia ‘a truly innovative, global, multilingual, collaborative knowledge-generating community and platform for authoring, editing, distributing, and versioning knowledge’ (Presner 2010). Larger than a mere technological innovation, wikis represent a change in the philosophy and practice of knowledge creation. Publishing scholarly work in such an environment is a direct intervention into multithreaded conversations maintained by lay knowledge communities on the web and existing scholarly discourses surrounding scholarly editing.

References

- Beal, Peter (ed.) 2008. *A Dictionary of English Manuscript Terminology, 1450-2000*. Oxford: Oxford University Press.
- The Devonshire MS Editorial Group. *A Social Edition of the Devonshire MS* (BL Add 17, 492). *Wikibooks*. http://en.wikibooks.org/wiki/The_Devonshire_Manuscript. Accessed 5 November 2014.
- DiNucci, Darcy. 1999. ‘Fragmented Future.’ *Print* 53. 32: 221-22.
- Leuf, Bo, and Ward Cunningham. 2001 *The Wiki Way*. Boston: Addison-Wesley Professional.
- Mahony, Simon. 2011. ‘Research Communities and Open Collaboration: The Example of the Digital Classicist Wiki.’ In *Digital Medievalist* 6.
- Marotti, Arthur F. 1993. ‘Manuscript, Print, and the English Renaissance Lyric.’ In *New Ways of Looking at Old Texts: Papers of the Renaissance English Text Society, 1958-1991*, edited by W. Speed Hill. 209-221, at 212. Binghamton: Center for Medieval and Early Renaissance Studies.
- Presner, Todd. 2010 ‘Digital Humanities 2.0: A Report on Knowledge.’ *Connexions* (Revised June 8 2010). http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6/Digital_Humanities_2.0:_A_Repo. Accessed 5 November 2014.
- Siemens, Ray *et al.* 2012. ‘Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media,’ *Literary and Linguistic Computing* 27. 4. 445-461. Oxford: Oxford University Press.
- Tonkin, Emma. 2005. ‘Making the Case for a Wiki,’ *Ariadne* 42. <http://www.ariadne.ac.uk/issue42/tonkin>.

Nowa Panorama Literatury Polskiej (New Panorama of Polish Literature)

How to present knowledge in the internet (Polish specifics of the issue)

Bartłomiej Szleszyński¹

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

Nowa Panorama Literatury Polskiej (New Panorama of Polish Literature, NPLP.PL) is a platform for the presentation of research results in the digital environment. Two first digital collections of NPLP.PL – *PrusPlus* and *The Literary Atlas of Polish Romanticism* – have been published on the 19th of September 2015. Currently, several other projects are being carried out by the NPLP team, some of which are going to be launched soon.

Scholarly base

NPLP.PL is a part of the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN), one of the most important Polish research institutions in the humanities. Digital collections in NPLP.PL benefit from a unique scholarly base and are meant to publish 'knowledge with the quality mark of the IBL PAN.' Adapting the form of presentation to the research conducted at the institute is one of the main goals of our work (Szleszyński *et al.* 2015).

¹ bartlomiej.szleszynski@ibl.waw.pl.

Structure

NPLN.PL does not have the structure of an encyclopedia. It consists of separate collections. Each collection is telling ‘a digital scholarly story’, using a different form to present content – each collection has its own narration, form, and visual identification. In the future, numerous collections will create a large knowledge base. The basic unit of meaning in NPLN.PL is an article with a normalized but flexible form: There are sections provided for elements such as graphics or maps, several functionalities for linking content or using hotspots on a map. There is also a special mechanism for publishing bibliographies and footnotes, as articles in NPLP are treated as pieces of scholarly knowledge. Each article can be searched using a built-in search engine.

Form

NPLP.PL tries to combine the high quality of published research results with an attractive, present-day, functional form, using the broad experience of NPLP team members in analyzing visual arts (digital, like video games or internet communication, but also more traditional formats like paintings and comic books). For each of the collections an individual form is created in order to communicate the specific content. This is why the interdisciplinary team of the New Panorama of Polish Literature includes not only literary and cultural studies researchers but also programmers, graphic designers and typographers. The interface is as intuitive as possible encouraging the user to immerse in the collection as deeply as possible.

The collection includes many visual materials – some of which are original graphics from the time period of each collection (e.g. in *PrusPlus* many graphics from 19th century newspapers from IBL PAN library were used); others are created especially for this collection.

Popularization of scholarly knowledge

Our activities are based on the general belief that there is a huge demand for high quality content on Polish literature. Since its launch NPLP.PL had over 27,000 unique users (and over 110,000 page views), which seems to confirm this assumption. NPLP.PL is also promoted via Facebook.² All collections are published in open access.

Digital editions and projects

Two digital collections have been published on NPLP.PL platform so far:

*PrusPlus*³ – a collection popularizing knowledge on Bolesław Prus and his most important novel, *The Doll*. It has a lexicon part, where articles are arranged by

2 <https://www.facebook.com/nowa.panorama.literatury.polskiej/>.

3 <http://nplp.pl/kolekcja/prus-plus/>.

different categories⁴, and a map-based part called *Warsaws of Prus*⁵ whose narrative is planned both to exploit the potential of stories which connect the writer and the space of the city and to debunk some untrue stereotypes, which have occurred over the years about Warsaw, Prus and *The Doll*.

*The Literary Atlas of Polish Romanticism*⁶ – a map-based collection which breaks with the traditional way of organizing knowledge on Polish Romanticism according to its authors. The project shows a different arrangement which is based on space. The collection contains almost 100 articles on different issues connecting authors and their work with space. Many interactive maps have been designed and created for this collection and are planned to be re-used for future collections on different topics and periods.

The IBL PAN has very a long and well-established tradition of scholarly editing. In October 2015, a five-year editorial project, *Trio from Skamander Group on Emigration*, started with the goal of editing 1500 letters of correspondences from Jan Lechoń, Kazimierz Wierzyński, and Mieczysław Grydzewski both in print and in digital format (applying a customized version of the TEI Simple schema). Another three-year research project has just started on the life and work of Henryk Sienkiewicz: *Postmodern Sienkiewicz: A Digital Laboratory*.

References

Szleszyński, Bartłomiej, Konrad Nicinsky and Agnieszka Kochanska. 2015. 'Jak przekazywać naukową wiedzę w Internecie. (Na marginesach kolekcji 'PrusPlus' w Nowej Panoramie Literatury Polskiej).' *Napis* 21: 348-359. Online: http://www.napis.edu.pl/pdf/Napis021_artykuly/NAPIS-2015_SERIA-XXI_s348-359_Bartlomiej-Szleszynski_Konrad-Nicinski_Agnieszka-Kochanska.pdf.

4 <http://NPLP.PL/prus-plus/kategorie-standardowe>.

5 <http://NPLP.PL/prus-plus/warszawy-prusa>.

6 <http://NPLP.PL/kolekcja/atlas-romantyzmu/>.

Digital Rockaby

Katerina Michalopoulou¹ & Antonis Touloumis²

Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.

Introduction

The object of our study was the construction of an artwork that transcribes and reformats the original written text of Samuel Beckett's *Rockaby* on the one hand, and creates a new experience of performing, reading, or watching it on the other. Before proceeding to the presentation and analysis of the new artwork, we shall briefly refer to the general philosophical context where its theoretical foundation lies.

According to Kittler the archive of knowledge and experience until the late 19th century was visually linguistically structured. The data of experience could be selected, stored and processed through the image and the writing. He claims that: 'Texts and scores – Europe had no other means of storing time. Both are based on a writing system whose time is (in Lacan's term) symbolic (...) all data flows, provided they really were streams of data, had to pass through the bottleneck of the signifier' i.e. they had to be translated in static signs that signify (Kittler 1999, 4). Finally according to Kittler, 'Writing stored writing – no more and no less.'

Things changed in the late 19th century due to the advent of the phonograph and the cinema. Kittler states:

What phonographs and cinematographs, whose names not coincidentally derive from writing, were able to store was time: time as a mixture of audio frequencies in the acoustic realm and as the movement of single-image sequences in the optical. Time determines the limit of all art, which first has to arrest the daily data flow in order to turn it into images or signs(...). To record the sound sequences of speech, literature has to arrest them in a system of 26 letters, thereby categorically excluding all noise sequences. (Kittler 1999, 3)

1 Katmichalopoulou@gmail.com.

2 Antouloumis@gmail.com.

Nowadays by using modern digital computing machines and new programming languages we no more reproduce images but we visualize routines. Generally we no longer construct models that imitate or represent the reality, but models that construct new realities, putting in question concepts traditionally given such as the notion of text, writing, performance, and reading.

So in the case of the transcription of Beckett's *Rockaby* the main issue for us was not to create a new shape of the text of the book but to design and visualize the rules that determine the evolution of the shape of the text through time. These rules may prescribe the reading process of the text too.

Short description of the original Beckett's 'Rockaby'

The play describes the procession to the death of an elderly woman. It is structured into four main parts. At first, we watch the woman's anxiety to live a bit longer and communicate with someone from the outside world. Gradually hope fades; the woman abandons the attempt being closed on herself and her place, remembers her mother's death and resigns herself to her own death in the very same way.

During the performance, we watch the woman sitting in a rocking chair and hear a taped voice describing the event. The text and the instructions given by the author for the direction of the play, prove its strict structure (Beckett 2006). Some of the elements that were recognized to create its dynamism are:

- Imperceptible changes in repetitive and rhythmic movements,
- Repetitions and light differentiations of words, sentences and the sets of sentences that they produce,
- Specific syntactic manipulations of language, and
- The use of voice recording.

Some of the key elements that create the rhythmic organization of the original work were analyzed and used in the new artwork. These are as follows:

- The metre of the monologue is visualized by the repeated movement of the rocking chair where the woman sits. Her movement is slow and steady. It is interrupted only at the end of each part and activated by request of the woman namely by the word MORE.
- The tempo of the rhythm, which in our case is considered to be the speed of the movement of the chair, is slow so as to get across a sense of ante mortem paralysis.
- Any action, any movement of the chair going back and fro, corresponds to the recitation of a sentence. Each sentence is structured rhythmically – with respect to the metre – through the sound of words, their meaning and their symbolism.

The new artwork

Through the new artwork, the mechanism of formation of the original text is attempted to be revealed – visualized. We designed a new mechanism, which proposes a new manner of reading, by using software which in turn is also concealed in the new artwork.

The structure of the original play text was analyzed through a series of diagrams. Diagrams as intermediates may prescribe without predefining, may define fields without describing properties. In this way the diagrams may become capable of leading to the creation of new spatiotemporal structures commensurate to the original ones. In this attempt to diagrammatize the text of the play we focused on its syntax.

In one of the diagrams created, the text was not read diachronically, according to the linear horizontal direction in which it was written. In contrast, it was studied synchronically, according to the vertical direction, which simultaneously penetrated the horizontal textual structure.

Thus, the lines turned into columns, the words of each line were classified into those columns according to the content or the sounds that they produced. The repeated words were removed from each column so that each word would appear only once. Very few words remained because the original text is produced by repetitions of few words and sounds. What resulted from this process is nine columns consisting of a few words, which, when combined in an appropriate way, can reconstruct the 250 sentences of the original text.

The new artwork was created by using a modern code written in Processing. Processing is an open source computer programming language built for the electronic arts, new media art, and visual design. The original text was encoded in Processing by the creation of an algorithm which transcribed the entire monologue phrases in some words. The nine columns were converted to nine concentric rings of different radius. If the Processing file is exported to JavaScript, the ‘Narrative’ is activated by the user. The interaction depends on their willingness to provide an input, which is a mouse click, in order to determine the outcome, which is the next sentence of the text. The rings rotate like gears do and stop only when a combination of words of different gears creates a sentence of the original text. The circular movement of the gears corresponds to the repetitive motion of the rocking chair in the play.

The relationship between the hand and the computer mouse provides dermal proximity with the technical picture of the computer. Our relationship with the image ceases to be more visual-optical and becomes haptic-tactile. The art work or animated image no longer lives outside of us, outside our sensory-perceptual system, but in proximity to it. It resides inside of us.

The correlation

- a. The neutral voice heard during Beckett’s *Rockaby* – the identity and origin of which is undefinable (is it the woman herself? is it a narrator of her story?), sets up the play ensuring its development in time. The narration is the motivating power. Although the voice heard is obviously taped, the narration is linked to

the present since the woman periodically activates it by pronouncing the word 'MORE'. In the new artwork, the computer user activates the next step of the narration by using the computer mouse. So, in a peculiar way he becomes a narrator too.

- b. In the original artwork, the apparent insertion of the rules of modern rhythmology by the writer marks his preoccupation with the rhythmic succession; it reveals his views in relation to the perception of space and time as composed by the dominant material of his art, which is the language mainly as sound and secondarily as meaning. For us the priority given by the author to the sounds was a research point for two main reasons:
- i. It reveals that the author composes his works as temporal structures.
 - ii. It reveals a key tool of composition, the sound patterns.

The new artwork respectively is also attempted to be designed as a temporal structure in which the animated image is considered to be the basic composing tool. We focused on incorporating in our structure the two basic theories of time – the A and B theories.

In theory A, which considers the time as based on a past – present – future axis, the present is dominant; so according to this theory one focuses only on the sentence created in the present time by the click of the computer mouse. This sentence is a present instant of a continuously transforming image.

Theory B focuses on the relative position of words, their potential participation in a sentence and mainly their constantly simultaneous presence (McTaggard 1908). What is important now is not the sentence produced but the words as an ensemble and their inner relations. The meanings are produced by the differences arising from the relations of the words, just as the meanings produced by the relations of the sounds of the words in the original work.

The whole, namely the sentence, is created as 'an order of proximity, in which the notion of proximity first of all has precisely an ordinal sense' (Deleuze 2004). The attempt to create such a structure in motion, where the words occupy relevant sites or positions, constitutes the major differentiation between the new artwork and Beckett's play.

Concurrently the new artwork is designed rhythmically as a set of sections or instants, where, as in the original artwork, priority is given to:

- The image, instead of the plot,
 - The motion, instead of the action,
 - The shadow of the self and the subconscious, instead of the character-hero.
- c. In the original work the speech becomes ambiguous through the relationship of the visual with the audial. That is, the image and the sound are autonomous. This strange relationship between the sequence of images and audio frequencies and the overall sense that they produce, that is to say this correlation between syntax and meaning, eventually leads to the creation of new meanings.

Since our perceptual mechanism is analogous to a recording device such as the magic notebook of Freud, the empirical data may exist only if it is inscribed in us.

The aforementioned ambiguity of the original play may be found in the new artwork due to the unlimited possible correlations of all the words of the text that co-exist on the computer screen.

Epilogue

Marshall McLuhan argues that the content of one medium is always other media: film and radio constitute the content of television; records and tapes constitute the content of radio; silent films and audiotape constitute that of cinema (McLuhan 1964). In the new artwork two languages appear and are used; both the language used by Beckett (writing and speaking), and the modern programming language in which the proposed mechanism is written. The programming language contains the previous as it uses the latin characters and their punctuation. The correlation of the two languages creates new signs.

Kittler argues that the dominance of a new media presupposes that something ceases not to write itself (1999). In our case, the computer screen and keyboard may be considered the new medium where the Processing may be inscribed, and consequently the composing mechanism of the new artwork may be visualized. A process hitherto not writable.

Moreover, the composing process is disclosed through the experience of reading by giving it an active character. Our eye cannot scan the text before reading it as it usually does. Every new sentence will arise only when our hand decides to move.

Stefane Mallarme claimed that literature is made up of no more and no less than twenty-six letters. In this Modernistic tradition, of which he was an initiator and Schoenberg one of the successors by his twelve-tone method of composition, fits the Beckett's *Rockaby* revealing the author's insistence on the material of words, the sounds. This is also the key element that the digital artwork has arisen as a composition of signifiers rather than signifieds or as a composition of forms rather than meanings.

By our work we propose a mechanism for a reformation of the Beckett's *Rockaby*. This mechanism may also enlighten the temporal design of the original text and constitute a tool for the creation of variants based on it. Hence the original text is considered to be a space – time element, the spatial layout of which can be translated in terms of temporal succession.

References

- Beckett, Samuel. 2006. *The Complete Dramatic Works*. London: Faber and Faber Limited.
- Deleuze, Gilles. 2004. 'How do we recognize Structuralism?' In *Desert Islands and other texts 1953-1974*, edited by David Lapoujade. New York: Semiotext(e).
- Kittler, Friedrich. 1999. *Gramophone, Film, Typewriter*, translated by Geoffrey Winthrop-Young and Michael Wutz. Stanford: Stanford University Press.
- McTaggart, J. M. E. 1908. 'The Unreality of Time.' In *Mind: A Quarterly Review of Psychology and Philosophy* 17: 456-473.
- McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. Canada: McGraw-Hill.

Appendix

1. W: **M o r e .**
V: till in the **END**
the day came
in the **END** came
close of a long day
when she said
to
h e r s e l f
whom else
time she stopped
10. ***time she stopped***
going to and fro
all eyes
all sides
high and low
for another
another like
h e r s e l f
another creature like
h e r s e l f
little like
going to and fro
20. all eyes
all sides
high and low
for another
till in the **END**
close of a long day
to
h e r s e l f
whom else
time she stopped
time she stopped

Figure 1: 1st Diagram by the authors, based on the original printed text.



Figure 2: 2nd Diagram by the authors based on the original text.

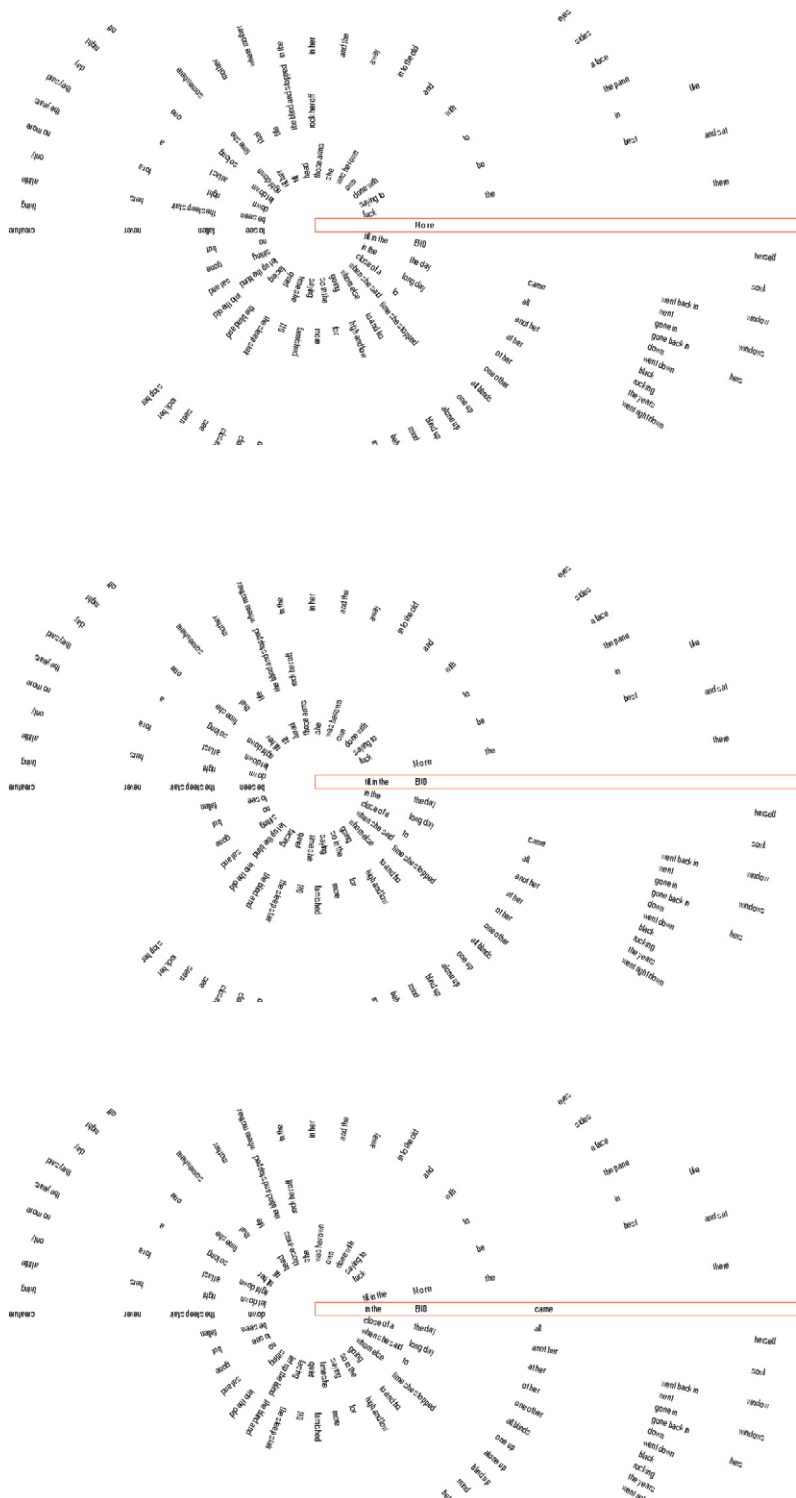


Figure 3: Stills from the animated text (Processing export).



PAPERS
PRESENTED AT
THE DIXIT
CONFERENCES
IN THE HAGUE,
COLOGNE,
AND ANTWERP

edited by
PETER BOOT
ANNA CAPPELLOTT
WOUT DILLEN
FRANZ FISCHER
AODHÁN KELLY
ANDREAS MERTGENS
ANNA-MARIA SICHANI
ELENA SPADINI
DIRK VAN HULLE

ADVANCES IN DIGITAL SCHOLARLY EDITING

As the papers in this volume testify, digital scholarly editing is a vibrant practice. Scholarly editing has a long-standing tradition in the humanities. It is of crucial importance within disciplines such as literary studies, philology, history, philosophy, library and information science, and bibliography. In fact, digital scholarly editing represents one of the longest traditions in the field of Digital Humanities — and the theories, concepts, and practices that were designed for editing in a digital environment have in turn deeply influenced the development of Digital Humanities as a discipline. By bringing together the extended abstracts from three conferences organised within the DiXiT project (2013-2017), this volume shows how digital scholarly editing is still developing and constantly redefining itself.

DiXiT (Digital Scholarly Editing Initial Training) is one of the most innovative training networks for a new generation of scholars in the field of digital scholarly editing, established by ten leading European institutions from academia, in close collaboration with the private sector and cultural heritage institutions, and funded under the EU's Marie Skłodowska-Curie Actions. The partners together represent a wide variety of technologies and approaches to European digital scholarly editing.

The extended abstracts of the convention contributions assembled in this volume showcase the multiplicity of subjects dealt with in and around the topics of digital editing: from issues of sustainability to changes in publication cultures, from the integrity of research and intellectual rights to mixed methods applied to digital editing — to name only a few.

Sidestone Press

ISBN: 978-90-8890-483-7



9 789088 904837 >

