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

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)

Introduction

Peter Boot,[1] *Franz Fischer*[2]

& Dirk Van Hulle[3]

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

The formalization of textual criticism

Bridging the gap between automated collation and edited critical texts

Gioele Barabucci[1] & Franz Fischer[2]

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.

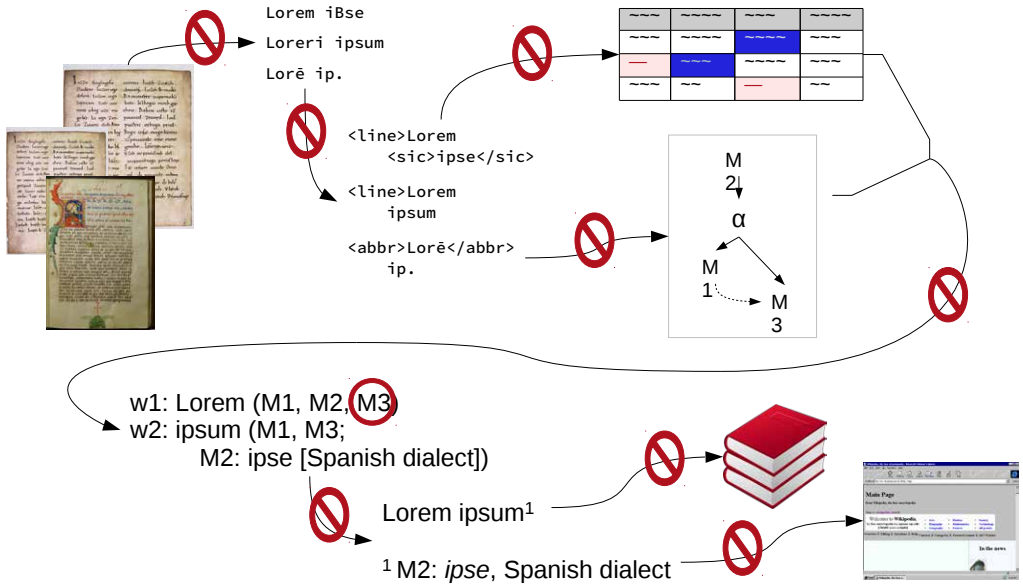

Figure 1: The problem: current tools do not communicate between each other.

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.

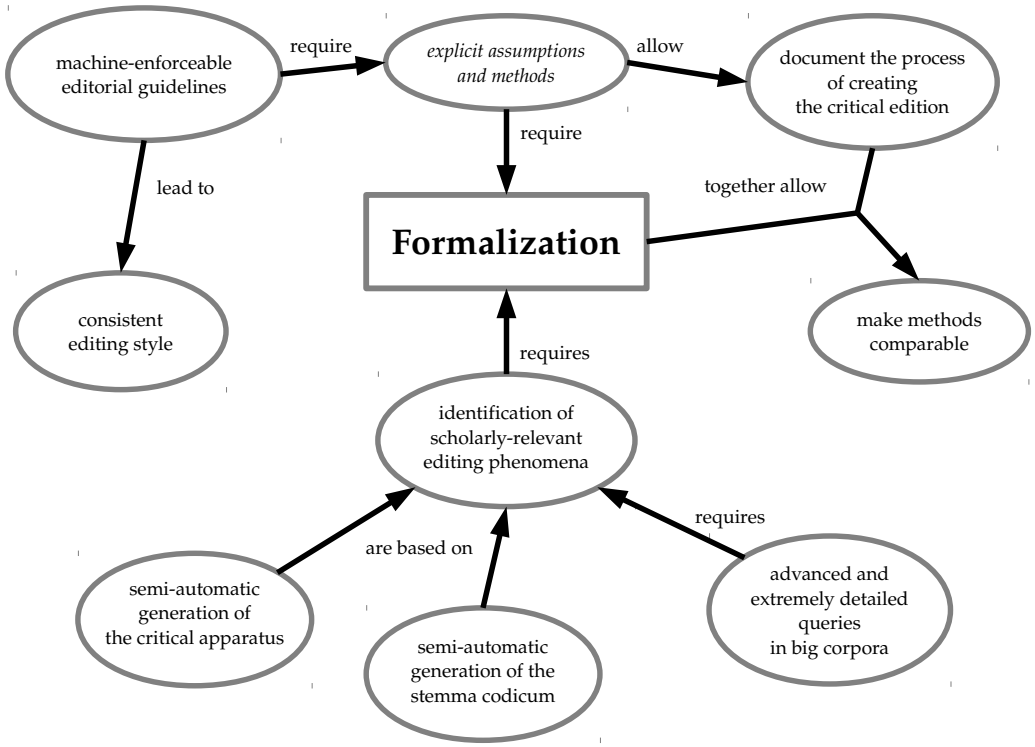

Figure 2: All the editorial steps need, directly or indirectly, a shared formalization.

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 ⇒ 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 paradigma 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.