Report of the Short Term Scientific Mission - COST action IS1310: *Reassembling the Republic of Letters, 1500-1800*

Elena Spadini

During the Short Term Scientific Mission within the COST action *Reassembling the Republic of Letters*, I analysed the tools and applications that support scholars in handling the transcription, annotation and collaborative editing of the texts of the letters. My contribution has been to pursue existing states and future prospects for editorial platforms, throughout an analytic and comparative study of transcription and annotation tools. Some of the ideas presented here will be tested during the Training school and design Sprint held in July in Como as part of the COST action.

In this Short Term Scientific Mission I worked in collaboration with Prof. Charles van den Heuvel at the Huygens Institute. This has been an excellent setting for this mission because I could benefit from the competences of researchers around me; various successful projects of text editing and text analysis have been completed at the Institute and others are underway. Furthermore, one of the tools that I studies, eLaborate, is developed at the Institute; this gave the opportunity to discuss the challenges that researchers and developers are facing in building it.

1. Editing tools and environments

What is a digital editing tool? It is, first of all, a program, a stand-alone or web application. Software is pervasive in the digital era, hence Manovich's notion of *software society* or *software culture*. In his words, 'software has become our interface to the world, to others, to our memory and our imagination' [Manovich 2013, p. 2]. Manovich's examples vary, ranging from hospital and transport management to email providers. In order to better understand what software is, we need to analyse 'its genealogy (where it comes from), its anatomy (interfaces and operations), and its practical and theoretical effects' [*ibid.*, p. 124]. This report will provide some targeted insight into these issues.

An editing tool is here defined as a piece of software used in the creation of a digital edition. Word processors, concordancers or lemmatizers are programs, i.e. digital tools; dictionaries and glossaries can be computer applications or appear in print, i.e. digital or non-digital tools; a set of paper forms for bibliographical resources is a non-digital tool. 'Extending the toolkit of traditional scholarship' [Burdick 2012, p. 8] is doubtless one of the aims of Digital Humanities. The core of each discipline (and profession) can be found in its toolkit and its applicability: a tailor might not anticipate the next dress to be commissioned by her client, but with the right measuring tape, thimble, shears (etc.), s/he can produce it. Within digital media, scholars have created digital equivalents to non-digital media. As Andrews points out, the question of digital publication has attracted a lot of attention; but 'the method of production, rather than the published form that the resulting edition take, is the practice wherein lies most of the promised revolution within textual scholarship' [Andrews 2013][1].

The development of ad hoc tools for the creation of a scholarly edition is not novel. The history of editing tools remains to be written, but some pioneering projects in the field are well known: TUSTEP, for instance: a toolbox for scholarly processing textual data, designed at the Computing

1 Cf. Pierazzo 2011: 'Perhaps we should just stop trying to map digital editions to printed ones and instead recognize that we are producing a different type of object, one that we can perhaps call a *documentary digital edition*. This new object necessarily comprises all three components of a digital publication—the source, the output and the tools to produce and display it—and it is worth emphasizing again that all three are scholarly products that result from editorial practice' (p. 474-5) and Van Zundert-Boot 2011: 'We argued that digital scholarly editions will be composites of three types of digital web services: data sources, processing services, and interfaces' (p. 150).

Center of the University of Tübingen, first implemented in the seventies and constantly upgraded until today; or Collate, a collation tool developed by Peter Robinson in the early nineties, which has only recently encountered a successor in CollateX [Dekker and Middell 2010-2014]. Theoretical reflection on digital tools for the humanities has been on the increase in past decades, fostered by landmark writings [e.g., McCarty 2005, Bradley 2002, Andrews 2013, Unsworth 2003], projects[2] and conferences[3]. On a practical note, a number of resources are available today; two main repositories of humanities applications, not focused only on editing but more generally on research tools for scholarly use, are DIRT (Digital Research Tools Directory)[4] and TAPOR (Research Tools for Textual Studies)[5]. The spreadsheet Collaborative Transcription Tools set by Ben Brumfield is a useful in-progress and collaborative resource[6].

The history of text-oriented applications swings between tools developed for specific uses and general-purpose ones[7], producing what has been called the 'tool paradox' [Pape, Schöch, Wegner 2012]: 'the conclusion would be that either “complexity is the price paid for generality” or that “limitation is the price paid for ease-of-use,” with both standing in the way of a widespread user base'. The authors continue: 'the challenge to combine expressive power with simplicity of use is faced by any developer in the domains of application software or operating systems, and it marks the long-term evolution of such endeavours'. The application of software development models to text analysis software development offers a promising path, but it does not seem to be a priority of humanists or even of digital humanists. This is changing; increasing interest is funnelled into, for example, user-centric design or web standards.

2. A comparative analysis

A number of editing tools and environments will be analysed here, focusing on encoding and transcribing while also considering collation for one of the environments. This selection follows a strictly empirical criterion of 'user-friendliness'. Only tools that require minimal computer literacy[8] and not to consult complex manual; only browser or portable applications, for which no installation is needed, have been selected[9]. Though this is a small selection of the available applications[10], it is worth engaging with these examples.

The tools taken into account are T-Pen and CWRC-Writer; the environments are eLaborate, TextGrid and Ecdosis. In the following table, the main features of these tools are summarized and compared.

2 E.g., Project Bamboo <https://wikihub.berkeley.edu/display/pbamboo/Documentation>, Interedition <http://www.interedition.eu/>.

3 Recently, *Easy Tools for Difficult Texts*, Cost Action IS1005 'Medioevo europeo' and Huygens ING, Den Haag, April 2013; *Research Summit on Collation of Ancient and Medieval Texts*, COST Action IS1005 'Medioevo Europeo', Münster, October 2014; *Scholarship in Software, Software as Scholarship: From Genesis to Peer Review*, Universität Bern, January 2015.

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

5 <http://www.tapor.ca/>.

6 <https://docs.google.com/spreadsheets/d/1MFsRSZRGy3RRB4AUD6AFp7IQsecqcauJLyZLGVzJFWs/>.

7 Unsworth distinguish four generations, from ad hoc programs, through reusable libraries of text-processing routines, to interactive general-purpose programs, which move over the network in the last stage [Unsworth 2003].

8 As said in the Preliminary, we include in the 'minimal computer literacy' basic knowledge of XML-TEI, as it is rather common in textual scholars who are interested in Digital Humanities and more specifically in producing SDE.

9 A tool that works in a browser is made up on web pages. Within an editor, the user interaction (typing of the text, pressing buttons, etc.) modifies the HTML source and the corresponding XML, if any, calling the proper JavaScript libraries. When the user gives the command, or automatically in time, the editor interacts with a server (or a web-worker) for saving and validating.

10 A lot of interesting projects falls outside this presentation, for instance Tustep <http://www.tustep.uni-tuebingen.de/tustep_eng.html>, the oXygen plug-in Ediarum <http://www.bbaw.de/en/telota/software/ediarum>, the Islandora TEI Editor <https://jtei.revues.org/790>, just to mention some of them.

3. Results

All the tools mentioned here are or include transcription and/or encoding facilities. The following table lists the common functionalities for transcription and encoding; other features are discussed below.

Object	Action
Text	add / edit / delete insert special characters copy / paste search save import / export
Image	add / edit / delete zoom in / out link text-image (image parsing or shape recognition, link mechanism, etc.) copy / paste save import / export
Metadata	add / edit / delete may be done through a form copy / paste search save import / export
Markup	add / edit / delete code completion copy / paste search save import / export

The tools enable different views of the content: XML (and XML structure), HTML, text, text and markup.

Certain applications, such as T-Pen, offer friendly functionalities for working with images; others,

like CWRC-W, are more oriented towards easy ways to markup and enrich the text.

In the case of a side-by-side view (image and text, annotated text and preview, normalised or diplomatic editions, etc.), synchronization on scroll-down is implemented.

The encoding structure may be free or fixed. In applications for archives, using for examples the EAD or CEI Guidelines, or in filling in a TeiHeader, the encoding structure is, more or less, fixed. A form can therefore be used to markup the text[11].

When there is no fixed encoding structure, it is more difficult to provide simple access to the high number of available tags, for instance TEI tags. The tool will propose the most common, using buttons or other intuitive mechanisms[12]; it is worth remembering that the more frequent tags do not correspond to the Bold, Italic and Underline buttons of most text editors: in a TEI editor, the tags <p> for paragraphs, <lb> for line breaks or <hi> for highlighted portion of text will probably be among the most used. The rest of the available tags may be accessible through menus or other graphical elements, referring to the modular structure of the TEI Guidelines[13]. Having tools that process the schema and customize the interface accordingly (for instance buttons and menus for accessing elements and attributes), may aid easier encoding mechanisms[14].

4. Correspondences, text encoding and linked open data

As already mentioned, when the encoding structure is fixed, it is possible to propose an interface as a form to fill in. This is normally the case for metadata, which would be the same over a series of documents within a project.

For letters, in particular, and other kind of correspondences materials (postcards, billets, etc.) the TEI implemented a new section of the Guidelines, proposed by the 'Correspondences Special Interest Group', and integrated in Spring 2015. The new element is called <correspDesc> and together with its subset of tags can be used to encode all metadata specific to this kind of texts.

A TEI encoding can be considered complementary, and not alternative, to other options. The Semantic Web approach and its technical standards, for instance, are becoming more and more central in structuring and representing cultural heritage data. A mixed use of XML/TEI and XML/RDF has been already implemented in few projects, among which:

- *Vespasiano da Bisticci. Lettere*.[15] In this case RDF is mainly used to build knowledge in the metadata section. Thus from occurrences inside the text, one points to the metadata using TEI; the metadata and the links between them (i.e. linked data) are then stored as triples in external RDF files, connecting to other projects and database as much as possible. Furthermore, an ad-hoc ontology has been built for this project, in order to handle the variety of relations among data.

- *Burckhardtsource*. Here RDF is directly based in the text. The starting point is similar, which is an XML/TEI version of the text. Then an HTML version is generated though XSLT transformation. Annotations (triples) are finally created, using the tool Pundit[16], that allows to annotate webpages.

11 Similar tools are LIME <http://lime.cirsfid.unibo.it/>, under development at the University of Bologna and mostly oriented towards the encoding of manuscript descriptions; and Doctored.js <http://holloway.co.nz/doctored/>, an application under development as well, which deal so far with the encoding of bibliographical references.

12 Cf. the customizable Author View of the XML editor oXygen.

13 See for example the developments at the Center for Textual Studies and Digital Humanities at the Loyola University Chicago <http://www.luc.edu/ctsdh/researchprojects/hrit-catt/>; demos are accessible at <http://hritwiki.ctsdh.luc.edu/demos>; the editor concept at <http://hritwiki.ctsdh.luc.edu/galleries/hrit-php-demos/tei-editor-concept>.

14 See for example Wed, an online schema-aware XML editor, developed at the Mangalam Research Center for Buddhist Languages <http://mangalam-research.github.io/wed/>.

15 See in particular 'La base di conoscenza', at <http://vespasianodabisticciletters.unibo.it/base_conoscenza.html>.

16 <https://thepund.it/> and <https://www.slideshare.net/netseven/iannotate-2016-demo-pundit-web-annotator>.

The stability of the URI (subject or object of the triple) rest of course upon the stability of the webpage.

Another aspect to be considered in handling resources for editions and analysis are the digital facsimiles, its metadata and the link between the image and the text or other resources. The Shared Canvas Data Model / IIIF offer support on this side, providing a conceptual template and technical implementation for a shareable and open integration of digital facsimiles.

Shared Canvas, together with the joint use of XML/TEI and XML/RDF, will be further explored during the Training School and Design Sprint held in July in Como, focusing on appropriate editorial interfaces for working on multiple correspondences at the same time.

5. Digital editing and NLP

The use of NLP (Natural Language Processing) is common in the process of creating a digital edition. Being NLP an extremely varied field, the majority of the editorial projects may use some NLP methodologies and tools within its workflow. We will briefly discuss some of them, which are widely spread.

When the source are printed texts, it is common to take advantage of OCR (Optical Character Recognition) for creating digital version that will serve as the base for further analysis, encoding, interpretation and editing. OCR outputs often need a human check; once common error are identified, they can be removed using regular expressions or other automatic process.

NER (Named Entity Recognition) is another methodology that is widely used in the field of digital editing and several tools, for different languages, are available on this side.

Part of Speech Tagging, Lemmatization and other syntactic approaches may be used in the early stage of a project where the linguistic aspect is particularly relevant. On the other hand, topic modelling, sentiment analysis and other methodologies for investigating the semantic layer, may be applied to the texts made available through the edition.

To summarize, NLP resources are often used in the pre-processing stage or after the edition is finished and published, but almost never integrated into the edition itself. This brings to the absence of documentation within the editions about if and how NLP resources have been used. Mostly in the case of semantic analysis, it is often not the editorial team, but someone else who used the edition data.

This state of the art seems to assess that

- the long-standing division between editorial work and literary analysis is valid also in the digital realm. In this sense, the editorial work is separated from (and may be ancillary to) eventual literary studies to be performed on the texts established by the edition;

- a digital edition is not (yet) considered as a set of data, mostly but not only textual data, that can be processed through a variety of algorithms and visualized in a variety of ways[17].

Not surprisingly, it is from outside editions of literary materials that some advances in the integration of digital editions and NLP are being made and experiments attempted. The ePistolarium project, collecting fundamental materials for the history of science, is at the fare-front of this process, using Topic Modelling, Keyword Analysis and other NLP methodologies.

17 See James Cummings, Martin Hadley, Howard Noble, ‘It has moving parts! Interactive visualisations in digital publications‘ in *The Educational and Social Impact of Digital Scholarly Editions*. DiXiT Workshop, Rome, 24 Januray 2017.

6. Works cited

Andrews, Tara. 2013. “The Third Way: Philology and Critical Edition in the Digital Age.” *Variants* 10: 61–76.

Andrews, Tara. 2014. “Digital Techniques for Critical Edition.” In *Armenian Philology in the Modern Era: From Manuscript to Digital Text*, Ed. V. Calzolari and M. E. Stone. Leiden: Brill.

Bradley, John. 2002. “Tools to Augment Scholarly Activity: An Architecture to Support Text Analysis.” In *Augmenting Comprehension Digital Tools and the History of Ideas*, edited by Harold Short, Dino Buzzetti, and Guiliano Pancaldi, 19–48.

Burdick, Anne et al. 2012. *Digital Humanities*. Cambridge, Mass: MIT Press.

Dekker, Ronald, and Gregor Middell. 2010-2014. *CollateX*. <http://collatex.net>.

Hockey, Susan. 2000. *Electronic Texts in the Humanities. Principles and Practice*. New York: Oxford University Press.

Leonardi, Lino. 2007. “Filologia Elettronica Tra Conservazione E Ricostruzione.” *Digital Philology and Medieval Texts*. Ed. Arianna Ciula, Francesco Stella. Ospedaletto (Pisa): Pacini.

Manovich, Lev. 2013. *Software Takes Command: Extending the Language of New Media*. London: Bloomsbury Publishing.

McCarty, Willard. 2005. *Humanities Computing*. Basingstoke [England]; New York: Palgrave Macmillan.

Pape, Sebastian, Christof Schöch, and Lutz Wegner. 2012. “TEICHI and the Tools Paradox.” *Journal of the Text Encoding Initiative* Issue 2.

Pierazzo, Elena. 2011. "A Rationale of Digital Documentary Editions." *Literary and Linguistic Computing*. 26.4: 463-477.

Pierazzo, Elena. 2014. “Digital Documentary Editions and the Others.” *Scholarly Editing: the Annual of the Association for Documentary Editing* 35. <http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html>.

Régnier, Philippe. 2014. “Ongoing challenges for digital critical editions.” *Digital Critical Editions*. Ed. Daniel Apollon, Claire Belisle, and Philippe Régnier. Urbana: University of Illinois Press.

Schmidt, Desmond. 2014. “Towards an Interoperable Digital Scholarly Edition.” *Journal of the Text Encoding Initiative* Issue 7.

Schmidt, Desmond, and Robert Colomb. 2009. “A Data Structure for Representing Multi-Version Texts Online.” *International Journal of Human-Computer Studies* 67.6: 497–514. <http://www.sciencedirect.com/science/article/pii/S1071581909000214>.

Sperberg-McQueen, Michael. 1996. Trip Report: Text Analysis Software Planning Meeting, Princeton, New Jersey. <http://www-01.sil.org/cellar/import/teilite/ceth9605.sgm>.

Unsworth, John. 2003. “Tool-Time, or ‘Haven’t We Been Here Already?’: Ten Years in Humanities Computing.” Washington, D.C.. <http://people.brandeis.edu/~unsworth/carnegie-ninch.03.html>.

Zundert van, Joris, and Peter Boot. 2011. “The Digital Edition 2.0 and the Digital Library: Services, Not Resources.” *Digitale Edition und Forschungsbibliothek*. (Beiträge der Fachtagung im Philosophicum der Universität Mainz am 13. und 14. Januar 2011) 44: pp. 141–152.