by way of a wiki so that dialogue about any letter, page, or single sentence could be documented.

Another crucial function of *Transkribus* is tagging and annotation. It is very easy to create new tags (that is, standardized, machine-readable labels to be attached to various parts of the transcribed text). In Tengnagel's case we have created, among others, tags for book titles, for languages such as Arabic, Greek, and Hebrew, for quotations, abbreviations, doubts in transcription, and so on, while tags for person names and place names already exist. *Transkribus* makes it possible to export any single page, group of pages, or entire document, in various formats (including PDF, DOCX, TEI, RTF). The user can therefore choose to export just the images, the transcription, or the tags.

However, what remains most exciting about *Transkribus* is the way in which its cumulative machine learning ability reshapes access to handwritten historical documents. Once a certain number of folios are transcribed, it is possible to train an HTR model and to perform an automated transcription of the texts. Every user benefits from the work of other users because the data are collected centrally, even though it is possible to keep each collection private, without the need to share documents directly. Within the Tengnagel project, a model has been created (using 5,478 transcribed words) that will be improved as the work continues. However, the project focuses more on enriching metadata by selective annotation; it will not go 'beyond' the reading process into digital editing, but rather think about metadata creation, pattern discovery, network analysis, and similar fields. Finally, since late 2017, the so-called Keyword Spotting feature has been included in *Transkribus*. This enables users to search for words directly in the image – and not only in the transcribed full text. This is a much more powerful method, and when recognition rates rise above 20–30 per cent CER, all words will be found with a high degree of confidence.

4 Producing Scholarly Digital Editions

Elena Spadini

In order to engage with the text of the letters of a correspondence, for distant as well as close reading and for producing a digital edition, the text must be available in electronic format. In the previous sections of this chapter, some of the methods for obtaining the text of a work in a suitable medium have been described, such as crowdsourcing transcription and automatic handwriting recognition. In this section, we focus on complementary strategies for producing scholarly editions in a digital framework.

A scholarly edition presents a reliable text and can take different shapes. In the case of a work transmitted by a single witness (that is, a manuscript, print, or born-digital copy), the text will be transcribed and normalized (more or less, from dip-

lomatic to interpretative), producing a documentary edition; this is mostly the case for correspondences. In the case of a work transmitted by various witnesses, a critical text is established, based on one witness or on the stemmatic reconstruction of the archetype, and accompanied by the variants from the other witnesses. A scholarly edition also provides materials useful for understanding the work's form and content, such as notes, introductions, apparatus, glossary, and others.

In the digital realm, a distinction must be made between digitized and digital editions. Digitized editions are the product of the digitization of an existing printed edition to make it available electronically, for instance, in PDF format. A digital edition, on the contrary, includes contents and functionalities (for example, advanced search, on-the-fly collation, interactive visualizations) that would be lost if converted to a printed medium.[46]

In what follows, we focus on different aspects of the creation of a scholarly digital edition. First, we define the term 'editing tools', which is the basic unit of an editorial platform, and provide a comparative study of a particular type of tool for transcription and encoding. Second, we explore the role of standards for encoding and annotation in the development of such tools. Finally, we briefly address the use of distant reading strategies and of Natural Language Processing (NLP) techniques in this context. The three sections pursue tasks that are connected, while potentially independent. Together, they provide insights into the process of creating a scholarly digital edition.

4.1 Editing Tools: A Comparative Analysis of Transcribing and Encoding Tools

An editing tool is here defined as a piece of software used in the creation of a scholarly digital edition: such a tool can potentially be used for transcription, annotation, collation, and, ideally, all the other tasks involved in the process.[47,48] A

[46] See Patrick Sahle, *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, vol. 3 (Norderstedt: Books on Demand, 2013), 141–2 and 149; Greta Franzini, 'Catalogue Digital Editions', https://dig-ed-cat.acdh.oeaw.ac.at/, accessed 20/03/2019.

[47] This section presents extracts from the DiXiT report 'Editing Tools. Transcribing and Encoding', including questions concerning the entire workflow necessary for the creation of a scholarly digital edition, see http://hdl.handle.net/20.500.11755/7227e906-2bad-4b8a-9611-1dd351d8bb85. A special issue of the journal *RIDE* (https://ride.i-d-e.de/) dedicated to tools and environments for digital scholarly editing is in preparation (November 2018); all accessed 20/03/2019.

[48] Considering several formalizations of the editing task and the peculiarities of a digital project, we can list: collection of witnesses (doc/image management and metadata), transcription, encoding, named-entity recognition, semantic enrichment, collation, analysis, constitution of the critical (or copy) text, compilation of apparatuses, compilation of indexes, preparation of paratextual material, data visualization. This list leaves out the final step of publication and might not include project-specific tasks. See for example Paul Maas, *Textkritik* (Leipzig: Teubner, 1927); Michael L. West, *Textual Criticism and Editorial Technique* (Stuttgart: Teubner, 1973); Tara Andrews, 'Digital Techniques for Critical Edition', in Valentina Calzolari and Michael E. Stone, eds., *Armenian Philology in the Modern Era: From Manuscript to Digital Text* (Leiden: Brill, 2014), 175–95, see https://doi.org/10.1163/9789004270961_008; Wilhelm Ott, 'Strategies and Tools for Textual Scholarship: The Tübingen System of Text Processing Programs (TUSTEP)', *Literary and Linguistic Computing* 15:1 (2000): 93–108,

number of editing tools can be combined in an editorial platform (also known as 'editing environment', or 'workbench'). Word processors, concordancers, or lemmatizers are programs, that is, digital tools. Dictionaries and glossaries can be computer applications or appear in print: that is, these tools may or may not be digital. 'Extending the toolkit of traditional scholarship'[49] is doubtless one of the aims of digital humanities. The core of each discipline can be found in its toolkit and its applicability: a tailor might not anticipate the next dress to be commissioned by her or his client, but with the right measuring tape, thimble, shears, and other tools it can be made. Within digital media, scholars have created digital equivalents of non-digital media. Furthermore, certain tools can only be available in electronic format. As Andrews points out, the revolutionary aspects of digital scholarly editing are not only related to the new publication media, but mostly consist in what happens, behind the scenes and behind the screen, in the process of editing with the aid of a partially new toolkit for textual scholarship.[50]

The development of ad hoc tools for the creation of a scholarly edition is not new. The history of editing tools remains to be written, but some pioneering projects in the field are well known. *TUSTEP*, for instance, is a toolbox for the scholarly processing of textual data, designed at the Computing Center of the University of Tübingen, first implemented in the 1970s and constantly upgraded until today. *Collate* is a collation tool developed by Peter Robinson in the early 1990s, which has only recently been superseded by its successor, *CollateX*.[51] Theoretical reflection on digital tools for the humanities has been on the increase in past decades, fostered by landmark writings,[52] projects[53] and conferences.[54] 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

see https://doi.org/10.1093/llc/15.1.93; Joris van Zundert and Peter Boot, 'The Digital Edition 2.0 and the Digital Library: Services, Not Resources', *Bibliothek und Wissenschaft* 44 (2011): 141–52.

[49] Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, *Digital Humanities* (Cambridge, MA: The MIT Press, 2012), 12.

[50] Tara Andrews, 'The Third Way: Philology and Critical Edition in the Digital Age', *Variants* 10 (2013): 61–76, see https://doi.org/10.1163/9789401209021_006.

[51] Ronald Dekker and Gregor Middell, *CollateX*, 2010, see http://collatex.net/, accessed 20/03/2019.

[52] See for example John Bradley, 'Tools to Augment Scholarly Activity: An Architecture to Support Text Analysis', in Harold Short, Dino Buzzetti, and Giuliano Pancaldi, eds., *Augmenting Comprehension Digital Tools and the History of Ideas* (London: Office for Humanities Communication, 2002), 19–48; John Unsworth, 'Tool-time, or "Haven't We Been Here Already?": Ten Years in Humanities Computing' (Washington, DC, 2003); Willard McCarty, *Humanities Computing* (Basingstoke [England] and New York: Palgrave Macmillan, 2005).

[53] E.g., 'Project Bamboo', see https://wikihub.berkeley.edu/display/pbamboo/Documentation; 'Interedition', see http://www.interedition.eu/, both accessed 20/03/2019.

[54] Recently, *Easy Tools for Difficult Texts*, Cost Action IS1005 'Medioevo europeo' and Huygens ING, The Hague, 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* (University of Bern, January 2015).

scholarly use, are *Digital Research Tools Directory* (DIRT)[55] and *Research Tools for Textual Studies* (TAPOR).[56]

Given the availability of so many tools for scholarly editing, a discussion of them all might quickly become overwhelming. A selection of them has therefore been analysed in the tables below, focusing on encoding and transcribing while also considering collation for one of the environments. This selection follows a strictly empirical criterion of 'user-friendliness'. This selection has been restricted to tools that require minimal computer literacy,[57] and to browser-based or portable applications, for which no installation is needed.[58] The tools that have been analysed are *T-Pen* and *CWRC-Writer*; the environments are *eLaborate*, *TextGrid*, and *Ecdosis*. All them have been tested[59] and compared. These tools and environments differ in so many different ways as to make a systematic comparison difficult. Some of their points of variation are represented in tables 1, 2, and 3 below.

Table 1: Comparative analysis of tools and environments (1)

	Web-based or standalone	**Licence**	**Documentation**
T-PEN	Web-based	Open source ECL-2.0	Users
CWRC	Web-based	Open source	Users
eLaborate 4	Web-based	Open source GNU GPLv3	Users and developers
TextGrid	Portable (connection with the server needed)	Open source Policy available on the website.	Users and developers
Ecdosis	Web-based	Open source	

[55] 'Digital Research Tools', http://dirtdirectory.org/, accessed 20/03/2019.

[56] 'Tapor', http://www.tapor.ca/, accessed 20/03/2019.

[57] Minimal computer literacy includes here basic knowledge of XML, but not of programming; also, the tools should be usable without consulting complex manuals.

[58] A browser application is a computer program which runs online, accessible through a website in the browser. *TextGrid* is the only software not running in the browser that is taken into account here; it is portable, in the sense that it does not require installation, but just to be copied and run.

[59] For an in-depth analysis, see note 47.

Table 1 describes technical features: all the selected tools and environments are web-based or portable, and have been released open source. Regarding the documentation, the solutions adopted vary greatly, and have consequences for ease of use by scholars and developers.

Table 2: Comparative analysis of tools and environments (2)

	Steps of the editing process	**Annotation and markup**	**Import and export**
T-PEN	Metadata management, transcription, encoding, annotation	Annotation. Possible TEI encoding	Import: TXT and XML Export: PDF, XML, plain text, HTML
CWRC	Metadata management, transcription, encoding, semantic enrichment	TEI encoding (*embedded*) and RDF encoding (*standoff*)	Import: plain text
eLaborate 4	Metadata management, transcription, annotation, publication	Annotation	Import: plain text Export: plain text, XML (not available on the normal user interface)
TextGrid	Metadata management, transcription, encoding, collation, lexicon creation, paratext creation, publication (+ Lemmatizer for German texts)	TEI encoding	Import: plain text, XML Export: plain text, XML
Ecdosis	Metadata management, transcription, encoding, paratext creation, event editor, publication	Markdown encoding	Import: plain text, XML Export: plain text, XML

In the first column of table 2, the distinction between tools and environments is evident, with the latter covering a higher number of steps of the editing process. The second column differentiates tools that allow for the encoding of the text from those offering annotation facilities, in combination or in alternative to the encoding. The import and export formats, mostly plain text or XML, detailed in column 3, are also dependent on the encoding choice.

The characteristics explored in table 3 concern the possibility to use these software packages as complete workbenches for the editing process: the image management, the search functionality, and the resources for online collaboration are important facilities common to the majority of the tools and environments under analysis.

Table 3: Comparative analysis of tools and environments (3)

	Search functionalities	**Image management**	**Online collaboration**
T-PEN	No	Advanced Text/Image link (line level)	Yes Leader project and users with different rights
CWRC	No	No	No
eLaborate 4	In metadata: advanced and *friendly* In text: full text search	Advanced Text/image link (page level)	Yes Leader project and users with different rights
TextGrid	Full text search in the project metadata and in TextGrid Repository	Advanced Text/image link (manual)	Yes Leader project and users with different rights
Ecdosis	No	Advanced Text/image link (word level)	Yes Leader project and users with different rights

4.2 Standards for Text Encoding and Annotation

The proliferation of transcribing and encoding tools depends to some degree on the acceptance of the TEI[60] as a standard for the encoding of texts.

Encoding is a fundamental step: a way of putting 'intelligence in the text'[61] – a kind of intelligence that computational tools can process. In practice, encoding in this context refers to the practice of inserting tags into the text, in order to label a portion of it (a letter, a word, a sentence or any combination of these) for the purpose of identifying or adding information. Typical examples include: the identification of a semantic entity, such as a name or a date; the identification of a structural entity, such as a title or a paragraph; or additional information, such as the expansion of an abbreviation or the comparison with other witnesses. The tags, or elements, labelling portions of text in this way constitute the 'markup'.

TEI provides Guidelines for this purpose and the data format it uses is XML.[62] XML-TEI is not the only existing data format for creating Scholarly Digital Editions: other XML languages, markup (such as LaTeX) and markdown syntax, or the code of web pages, HTML, can be used. However, XML-TEI markup, extremely rich and explicitly devoted to text encoding, has become a standard for transcribing and editing literary texts and historical sources, which are normally the objects of scholarly editions. The TEI Guidelines provide around 500 tags for marking up all sorts of phenomena occurring in texts. A quick look at the Guidelines Table of Contents will suffice: the twenty-three chapters are devoted each to a different aspect of text encoding, including verse, performance texts, dictionaries, language corpora, writing modes, linking, and many others.

The use of TEI as a standard plays an important role in the development of tools: a TEI-compliant tool is useful for a large community of practitioners,[63] and this is one of the reasons for the production of increasing numbers of transcribing and encoding tools that use this standard.

Using TEI for encoding correspondence material (letters, postcards, billets, etc.) is equally common. The Guidelines provide specific tags for the markup of the text and for the corresponding metadata (that is information about the document such as sender, receiver, date).

For the text, in particular, the tags for encoding default structures are available (TEI Guidelines, ch. 4: Default Text Structure); some of the tags frequently used

---

60 'Text Encoding Initiative', see http://www.tei-c.org/, accessed 20/03/2019.

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

62 The history of the Text Encoding Initiative is deeply intertwined with that of humanities computing (or digital humanities) and of the XML specification. See 'TEI: History', see http://www.tei-c.org/About/history.xml, accessed 20/03/2019.

63 As the TEI Guidelines are extremely rich and, more importantly, the object to encode, that is – text – is equally variegated, a customization of the TEI is likely to be used in each specific project. This makes it difficult to provide tools that can work out of the box for every single project, without some degree of customization.

for letters are <opener>, <closer>, <dateline>, <salute>, <signed>, <postscript>.[64]

For metadata, the Correspondences Special Interest Group[65] proposed a new section, which was integrated in the TEI Guidelines in Spring 2015. The element carrying the correspondence metadata is <correspDesc> (correspondence description).[66] Together with its subset of tags, it can be used to encode information about the sending, the receipt, the transmission, the redirection, and the forwarding of a message.

Using the Correspondence Description model as an interoperable standard for the encoding of correspondence metadata, the research group TELOTA has promoted the project *correspSearch*, with the aim of indexing letter collections. The database, which can be queried on the project website, already hosts more than 27,000 letters.[67]

Before the addition of the Correspondence Description section to the TEI Guidelines, one of the most successful projects in the realm of XML encoding of correspondence has been the *Digital Archive of Letters in Flanders* (DALF)[68] created by the Centrum voor Teksteditie en Bronnenstudie. The schema designed is an extension of the TEI P4 DTD,[69] and has been adopted with modifications in other projects, as the well-known edition of Van Gogh's letters.[70]

A TEI encoding can also be considered complementary, and not alternative, to other forms of annotations. The Semantic Web and its technical standards, for instance, are increasingly central in structuring and representing cultural heritage data. A mixed use of XML-TEI and XML-RDF has been already implemented in a few projects devoted to correspondences, among which are *Vespasiano da Bisticci. Lettere*[71] and *Burckhardtsource*.[72]

---

[64] The definitions of these elements, together with technical information and examples, are part of the TEI Guidelines. In particular, 'Elements Common to All Divisions', see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSDTB, accessed 20/03/2019.

[65] Information on Special Interest Groups is available at 'TEI: SIGs', see http://www.tei-c.org/activities/sig/, accessed 20/03/2019.

[66] 'Correspondence Description', see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD44CD, accessed 20/03/2019.

[67] 'correspSearch', see http://correspsearch.net/index.xql, accessed 20/03/2019.

[68] Centrum voor Teksteditie en Bronnenstudie. 'DALF', see http://ctb.kantl.be/project/dalf/, accessed 20/03/2019.

[69] An XML schema defines which tags, where, and how they should be used in an XML document. The DTD format for schemas has now been replaced by RNG. A TEI customization is expressed through an XML/ODD document. P4 refers to a previous release of the TEI Guidelines, the current one being P5.

[70] Jansen Leo, Hans Luijten, and Nienke Bakker, eds., *Vincent van Gogh: The Letters* (Amsterdam and The Hague: Van Gogh Museum & Huygens ING, 2009), see http://vangoghletters.org, accessed 20/03/2019.

[71] See in particular 'La base di conoscenza', at 'Vespasiano da Bisticci, Lettere', see http://vespasianodabisticciletters.unibo.it/base_conoscenza.html, accessed 20/03/2019. In this case, RDF triples are mainly used to build knowledge in the metadata. From occurrences inside the text, a link is established to the metadata, using TEI; the metadata and the relations between them (i.e. Linked Data) are then stored as triples in external RDF files, connecting to others data sets. An ad

4.3 NLP and Editions as Data

In a scholarly digital edition, the images of the documents, texts, and additional materials (notes, explanations, introductions) are not organized in a book, but digitally: all the information becomes electronic data, which is organized in an information architecture.[73]

Once the scholars engage with data, new forms of analysis become possible. Text in digital form is data, to which NLP techniques or Data Mining algorithms can be applied. Furthermore, data is structured: every individual structure can be statistically inspected and visualized. A closer look at each of these possibilities will clarify the potential of this approach.

NLP is a comprehensive label, referring to the processing and analysis of text and speech by means of computer applications. Among the techniques that are most used in the field of scholarly editing, we can find algorithms for linguistic annotation (such as part-of-speech tagging or stemming) and for Named Entity Recognition (NER). While the former adds information to the word, automatically annexing the corresponding part-of-speech or stem, the latter identifies the proper nouns in a text. NER can also be used for automatically encoding semantic entities such as names of persons and places (for instance, in TEI, see above).

NLP techniques might indeed be useful in the process of creating an edition, just as much as linguistic competences have always been necessary for editing a text. Data Mining, on the other hand, would mostly be used at the end of the process, exploiting the edition as a *corpus* for investigation. Data Mining can either take the form of quantitative analysis leading to data visualization or consist of the application of statistical models, as in the case of topic modelling, stylometry, or sentiment analysis.

A look at the current state of the digital editing field shows that projects seldom embrace the approach just described, taking advantage of the (generally) large amount of data produced in scholarly editing. In most of the cases, the editorial work, the linguistic analysis, and the text analysis remain separate, reproducing the traditional distinction between textual criticism and literary criticism. NLP and Data Mining techniques are not admitted into the editing process; thus their results cannot be integrated in the auxiliary materials provided alongside the text in a scholarly edition.

This state of the art might have multiple explanations. Among these is the fact that for a long time, producing scholarly editions has been considered an ancillary

---

hoc ontology has been built, in order to handle the variety of relations among data specific to the project.

72 'Burckhardt Source Project', see http://burckhardtsource.org/, accessed 20/03/2019. In this edition, an HTML version is generated from the TEI-XML encoding, following common practices. Annotations (triples) are finally created, using the tool *Pundit* that allows annotation of web pages.

73 An information architecture is simply some structure for organizing the data, for instance, the schema for the markup (see above section 4.2), the ensemble of database fields or the ontology for Linked Data.

activity in relation to historical and literary studies (*philologia ancilla historiae*), ratifying their separation of concerns. Most importantly, in the context of digital editing, an edition has not yet come to be regarded as a set of data that can be processed by means of a variety of algorithms and visualized in different ways: in the words of Cummings and colleagues, 'Many editors often think they have "only text", not "data"; but the structures created by the edition contain more data than they think'.[74]

This 'misunderstanding' of the nature of text and data and of what can be done with them appears to be stronger whenever literary materials are involved: it seems less pronounced for the edition and analysis of historical, and more generally cultural, sources. Indeed, it is in the edition of correspondences that some advances in the integration of scholarly editing, NLP, and Data Mining are to be found. The *ePistolarium* project,[75] providing the edition of letters of scholars active in the Netherlands in the seventeenth century, is at the forefront of this process. Using techniques such as topic modelling and keyword analysis in combination with NLP, it offers meaningful ways to engage with the text of correspondences. As such, it serves as an inspiring example for imagining the digital scholarly editions of the future.

5 Digital Provenance of Texts and Additional Resources: EMED

Elizabeth R. Williamson

Reassembling the republic of letters requires large-scale collaboration on previously unedited manuscript sources, but no less important is the repurposing of existing print and digital materials which have already absorbed vast amounts of scholarly labour. More generally still, any collaboratively compiled catalogue, archive, or edition builds on the work of many scholars, sometimes spread over several generations. This section will suggest how much care and critical attention will be required to ensure that due weight is given to the layered histories, agencies, and labour buried within any collaboratively crafted resource of this kind.

Digital editing itself has the potential to be a highly collaborative and highly iterative process: not only can many people work simultaneously or successively on a digital edition, but digital data in standardized formats lends itself to reuse by further groups with divergent purposes at different times. This potential introduces complexities around provenance, attribution, and acknowledgement that are rele-

[74] James Cummings, Martin Hadley, and Howard Noble, 'It Has Moving Parts! Interactive Visualisations in Digital Publications', paper presented at the DiXiT Workshop 'The Educational and Social Impact of Digital Scholarly Editions', Rome, 2017. See http://dixit.uni-koeln.de/programme/materials/#aiucd2017, accessed 20/03/2019.

[75] *ePistolarium*, see http://ckcc.huygens.knaw.nl/epistolarium/, accessed 20/03/2019.