Schriften des Instituts für Dokumentologie und Editorik — Band 16

Digitale Edition in Österreich
Digital Scholarly Edition in Austria

herausgegeben von | edited by

Roman Bleier, Helmut W. Klug

2023

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 29. April 2023.

2023

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-743-102-842
Einbandgestaltung: Stefan Dumont; Coverbild: wurde von Roman Bleier und Helmut Klug für ein KONDE-Poster (DHa 2017) erstellt
Satz: Roman Bleier und LuaTEX

Where are the Tools? The Landscape of Semi-Automated Text Edition

Tara L. Andrews

Abstract

The aim of this article is to answer the question: given that there have been so many tools and methods developed to help prepare scholarly critical editions of texts, why do so many scholars have trouble knowing where to start? The article walks the reader through the typical process of creating an edition, mentioning along the way a variety of tools that have been developed or used in the Austrian landscape in particular, and aims thereby to illustrate many of the considerations that the scholar setting out on an edition project must account for.

Zusammenfassung

Ziel dieses Artikels ist es, eine Antwort auf die Frage anzubieten: Wenn so viele Werkzeuge und Methoden entwickelt worden sind, um kritische Editionen von Texten vorzubereiten, warum finden es dann so viele WissenschafterInnen schwierig zu wissen, wo sie anfangen sollen? Der Artikel führt den Leser durch den typischen Prozess der Erstellung einer Edition, erwähnt dabei eine Vielzahl von Werkzeugen, die insbesondere in der österreichischen Landschaft entwickelt wurden oder verwendet werden, und versucht damit viele der Überlegungen sichtbar zu machen, die zu Beginn eines Editionsprojekts berücksichtigt werden müssen.

As more or less any professor, teaching fellow, or research assistant in the field of Digital Humanities can attest, there is a great deal of interest from scholars in the literary and historical fields about digital editions of texts and how they might be feasibly done. Yet many of these scholars have little idea where to start or what tools exist that are relevant for the particular work they wish to do, despite the fact that textual criticism has been on an increasingly digital trajectory since well before the availability of the World Wide Web. After so many years of development of the digital edition, this seems like a rather odd state of affairs—there are, after all, a plethora of tools available for use (Klug, Galka and Steiner 2021; see also the discussion of several of these tools and methods in Vogeler 2019). Why, then, can it be so difficult to advise scholars about a way forward?

The purpose of this article is to walk the reader through the process of creating a digital edition, from initial transcription to publication and including various methods

Digitale Edition in Österreich. Hrsg. Roman Bleier und Helmut W. Klug. Schriften des Instituts für Dokumentologie und Editorik 16. Books on Demand, 2023, 3–17.

of source analysis. Along the way we will cover a range of tools that have emerged, especially in Austria, to assist with the creation of digital editions. We must, however, stress the following. Although many textual scholars hope for a single full-featured software package designed to take their editions all the way from conception to publication without the need either to learn their way around computer programming or to hire someone who does, this hope is sadly misplaced. There are almost as many possible forms of digital edition as there are texts to be edited. Every editor will have a different set of priorities for her edition, not to mention a text (or a corpus) that differs from other texts in ways that are perhaps small, but certainly crucial, for the purpose of preparing that edition. While the argument has been made elsewhere that there can be no "monolithic" general-purpose tool (van Zundert and Boot 2011), we hope with this walk-through of digital editing processes to illuminate why this is the case.

Given the high degree of specialization that any edition project must reach, a scholar who wishes to produce a digital edition should expect to exercise a significant amount of control over the process. As such, the scholar will need to know the principles, and the limitations, of the data modelling system that is used to render texts into the digital medium, and will need to understand how and where the choices made for her particular project might differ from the assumptions built into the tools that are available for analysis and publication of the result. In many cases, in fact, different tools take a different set of assumptions as their starting point, and so the editor will eventually need to understand the data models and their associated technologies well enough to assess how—or indeed whether—these differences can be bridged. We contend here that, in order to create a digital edition on time and within budget, a scholar cannot hope to rely entirely on "IT experts" hired for the purpose. She will need to gain enough knowledge, not only about how text encoding is done, but also about what is done with the result of that encoding and the parameters of the technologies that are used to do it, to be able to make informed decisions.

1 From scholarly work to digital model

Texts can be published into the digital medium by a variety of means. The basic requirement for any online publication is that it must be expressed as one or more documents rendered in HTML—the standard format for web documents—and hosted on a server connected to the Internet, under a publicly reachable URL. In order for this publication to be a critical edition, it is only necessary that those documents, in one way or another, contain a faithful representation of the critical text and any apparatuses or other commentary that the editor felt necessary to include.

There are many possible pathways to this end state. For example, it is easy to envision the preparation of the edition in a word processor, where the document is then saved into HTML format and given to a hosting provider. That would result in an online publication that may be "digital enough" for many purposes, but could not be considered "a digital edition" in the sense proposed by Sahle (2016).

A middle ground between the print and the digital can be found with software such as the *Classical Text Editor* (CTE), developed at the Austrian Academy of Sciences (ÖAW) by Stefan Hagel (1997-). This is a package intended specifically for the creation of critical editions from multiple witness copies. Its user interface is intended to hew as closely as possible to the familiar interface of a word processor, while extending the functionality to provide for the things that an editor will need, such as the definition of sigla to correspond to witnesses, the possibility to record variant readings to the edited text based on those witnesses, the possibility to add other sorts of scholarly apparatus according to the conventions of classical philology.

Although the primary output of CTE is a print-ready document suitable for submission to a book publisher, it also offers the option of export to a TEI-XML format, which could then be transformed to HTML and published online (see below). This feature, added by CTE's author in the hope of encouraging the proliferation of companion digital publications by CTE users alongside the more usual print publications, has not had widespread use (Hagel 2007, 78). The author attributes this to a lack of institutional interest in digital publications; while this may still have been true in 2007, it is much less true today, and yet those who call themselves digital philologists still do not usually recommend CTE as a means to produce a digital edition. The reasons for this, we would argue, go deeper than institutional interest.

One of the major design decisions of CTE is to allow the scholar to create a text edition that conforms to her exacting scholarly specifications, but without troubling her with "any sort of surfacing tags or other sorts of 'code' [which] can be detrimental" to the ability of the editor to "remain devoted to scholarly questions" (Hagel 2007, 79). While this is an admirable goal, it may actually be part of the problem. CTE encourages the scholar to concentrate primarily on how the edition will look once it is printed. Although the software has a reasonably complex conceptual model, the scholar not inclined to study this model or to familiarise herself with the advanced features of CTE will quickly find that she can produce an edition that "looks right" on the page even when the model is violated internally.

For example, CTE provides a mechanism for recording additions, omissions, and transpositions, and for specifying the abbreviations that should appear in the *apparatus criticus* when one of these situations arises. The user can, on the other hand, achieve the same outward effect simply by inputting these abbreviations as though the abbreviation was itself the text of a variant. The distinction is invisible in the resulting print proof, but in the companion XML output, incorrect use of the data model will instantly undermine any automated attempts to parse the document for the edition text and the variant witness texts. Several potential inconsistencies of this type were encountered in the attempt to write a parser for CTE's flavour of TEI-XML, for use with the Stemmaweb service (Andrews n.d.).

It is thus clear that, if a user of CTE wishes to produce a *useful* digital output alongside the print output of an edition, she must take care to understand not only the scholarly features of CTE, but also the data model that underlies what is displayed on the screen. She will soon find that there is a set of technical assumptions embedded in the software about how to represent the data that has been entered by the scholar, and more assumptions about how to translate the text from the CTE model into the TEI double-endpoint-attachment means of encoding variant text. She will then need to draw upon her understanding of the data model as it was expressed in the TEI output, in order to use the tool(s) that she eventually chooses for further analysis of the edited text or for its publication.

2 The typical digital edition

Although the XML export functionality of CTE offers a route to HTML-based online presentation, this might not necessarily be considered a true digital edition. Many scholars follow the definition of "digital edition" offered by Sahle (2016), who argues that "scholarly digital editions are scholarly editions that are guided by a digital paradigm in their theory, method and practice"—for Sahle this also means that a digital edition must offer some significant content and functionality that would be lost in an analogous print edition.

What does this mean in practice? Although Bordalejo (2018) argues, with some justification, that textual scholarship has not actually progressed beyond traditional paradigms on a methodological level, a more or less mainstream set of steps have emerged toward creating something that is commonly acknowledged as a "digital scholarly edition". It begins with the transcription of sources, usually from digitized images, in which not only the sequence of interpreted text but also the features of the textual expression (such as decorated or otherwise highlighted text, authorial or scribal corrections, marginal notes, and the like) are recorded. The ability to pay close attention to these features of the text brings the scholar immediately beyond the capabilities of word processors, or even the CTE.

Thus, when we speak of a "typical digital edition", we usually envision the transcription of source text from digitized images, the comparison (if necessary) and analysis (as desired) of these source texts, the production of a commentary usually including information that is linked to specific portions of the text or specific words therein, and the presentation of the whole in an online format, so that a reader (or user) of the edition can view the text and its commentary in whatever form best suits the purpose that the editor had in doing the work.

Several institutions within Austria—the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) at the Austrian Academy of Sciences, the Centre for Information Modelling—Austrian Centre for Digital Humanities (ZIM-ACDH) at the University of Graz, and the Digitalisierung und elektronische Archivierung (DEA) group at the University of Innsbruck foremost among them—have taken on the task of supporting Austrian researchers by providing training for scholars new to digital methods, offering archiving of the digital sources, producing systems for transcription of digitized images, and providing solutions for publication of the finished result on the Web. Edition projects and the expertise that goes with them are settled at many other Austrian universities, and these are often an early port of call for the scholar who would begin a new digital edition.

The plethora of services around digital editions in Austria brings us to a first important point: although there exist a great variety of tools for various steps in the process of creating a digital edition, there will almost certainly not be a complete suite of tools that will take the scholar from the beginning of her project all the way through to publication, without the intervention of a programmer hired for the purpose or a specialist consultant, unless the scholar is herself conversant enough with the various relevant technologies to provide her own technical support. We can illustrate this point by considering the steps toward the production of such a “typical” edition, reviewing what tools are most helpful for each of these steps, and observing what is still missing.

3 Transcription as an act of data modelling

The first stage of a digital edition is to express the content of each source document in some sort of digital model. This expression usually takes the form of *embedded markup*, in which semantic information about the text is entered directly alongside the text itself. Although there are a number of options for embedded markup, such as LMNL (Piez 2014) or a system currently in development known as TagML (Haentjens Dekker et al. 2018), by far the most predominant toolkit for the expression of this model is the Text Encoding Initiative (TEI). The first step for the vast majority of trainee digital philologists is to become familiar with the TEI Guidelines, and how they are expressed in XML. We must stop short of calling the TEI itself a model; although it defines a vast number of scholarly concepts related to text and its features, many of these concepts are intentionally flexible in their definitions (e.g. text block definitions such as `<ab>` and `<div>`), and there are often multiple possible ways to express the same textual feature (e.g. lines of text, which can either be enclosed in a `<line>` element or separated by an `<lb/>` [line break] milestone element). This is the reason that the authors of the TEI Guidelines advise all editors to produce a custom schema, specific to the needs of the project itself, at its outset.

Since the Text Encoding Initiative has based its work almost entirely on the framework of XML and its related technologies, the first consequence for the newly-digital philologist is that, from the very outset of the transcription work, she must learn a great deal about XML. This includes not only the basics of its grammar, but also the subtleties of how that grammar is employed via the TEI Guidelines to create a document that would be considered "valid". Here too, the editor must understand the technical mechanism by which validity is ensured: this involves the creation and configuration of a custom XML schema using a tool such as Roma (Mittelbach, Rahtz, and Bernevig 2018), for which the editor will need to understand the technical contents of the TEI modules that she wishes to use. If the editor wishes to extend the TEI to deal with features of her text that are not adequately covered in the Guidelines, she will need to have an even better understanding of the principles of schema description in order to add or modify the required elements, attributes, or dependencies.

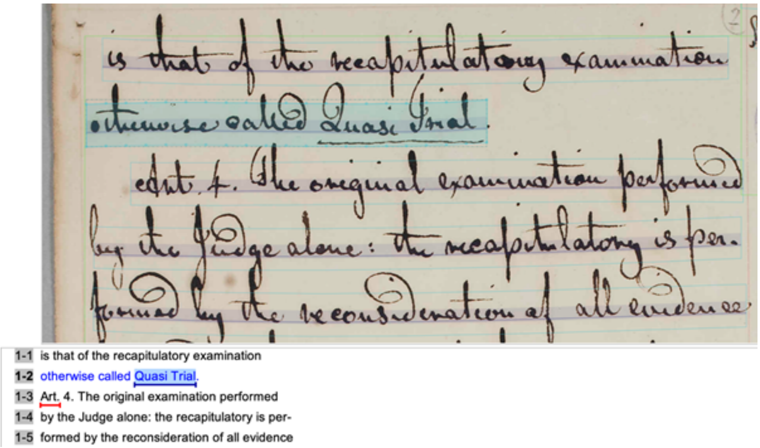
In many cases, the editor will do the transcription using an XML editor (the most commonly used editor is *oXygen*, though there are open-source alternatives) configured to incorporate her custom schema for validation checking. The very fact that XML editors are the primary tool of choice for creating TEI-XML transcriptions of source texts constitutes strong evidence of the impossibility of providing the comprehensive and user-friendly software tools that scholars so often wish for. Industry programmers, as well as colleagues from the field of computer science, are (in our experience) almost invariably stunned to discover that digital philologists write XML directly in an editor—although XML is a text format that is comprehensible by humans, its syntax is exacting enough that software developers almost never write it directly if they can avoid it. Rather, they expect that data in XML format is generated by some sort of intermediate software, and only edited by hand *in extremis*. This manual process is, however, almost inevitable when the very schema against which the transcription is checked varies from project to project. It is not uncommon for user-friendly alternatives to be developed within the bounds of individual projects—both the *Transcribe Bentham* crowdsourcing project (Causer and Wallace 2012) and the *New Testament Virtual Manuscript Room* (Institut für neutestamentliche Textforschung n.d.) implement WYSIWYG-style transcription interfaces, for example—but none of these interfaces makes a claim to widespread utility.

We say it is "almost inevitable" that the editor will write XML by hand—for the scholar who has facsimile images of her sources and wishes to transcribe them line by line with a view to publication of both facsimile and transcription, there are tools available that handle transcription of the text, together with line-by-line linking of transcriptions to facsimile images. One of the best-known tools for this, particularly in Austria, is *Transkribus* (Kahle et al. 2017). *Transkribus* is offered as a desktop application, backed by a central data storage and processing service, for transcription of manuscripts and printed texts directly from images. It offers a plethora of useful tools; these include automated detection of regions of text in an image and lines within those regions (which allows for the association of segments of transcribed text with their corresponding places on the facsimile image), handwritten text recognition (if enough of a particular document has been manually transcribed), and output of the transcription data into several formats, including TEI-XML.

Given the claim above that it is more or less impossible to write a graphical user interface for production of TEI-XML encoded texts, we should look more closely into what *Transkribus* does offer. Their data model, tuned as it is for recognition of text blocks on page facsimiles and association of text with these blocks, saves information using a standard known as PAGE XML (Pletschacher and Antonacopoulos 2010) that was developed for automated document analysis and text recognition. Any conversion to TEI must therefore involve a transformation of the model of a text as conceived by PAGE into some TEI-compatible model. Indeed, in their introductory How-To guide, the developers note that "Transkribus is [...] more than a TEI editor, but also less (we will not support all peculiarities of TEI but just those which are necessary to create a good, standardized transcription)" ("How to Use Transkribus—in 10 Steps (or Less)" 2015).

What this means in practice is that the commonly recommended option of creating a custom XML schema for each individual project is not an option here. The *Transkribus* developers have had to make certain decisions of their own about how the PAGE XML model can be expressed in terms of TEI-XML concepts, and the user has little choice but to learn and understand the respective models if she wishes to use the TEI-XML output of *Transkribus*. For instance, the PAGE XML concept of a TextRegion (that is, the area on the facsimile that comprises a text block, or a line inside a block) demands that the content of the line be nested inside this element. This in turn disrupts the usual use of text-structural tags such as `<p>` (paragraph), `<q>` (quotation), or `<head>` (heading), which now cannot be used for text that spans multiple lines without violating the strict hierarchical principle of XML. Even if the user chooses to export a version of TEI-XML that employs `<lb/>` (line beginning) milestone tags instead of `<line>` tags containing the line content, she will find that no markup has been allowed to cross a line boundary.

The developers, perhaps sharing the ideals of Stefan Hagel in wanting to ease the complexity of the data model and its expression as far as possible, have pre-configured a few markup tags for the convenience of the user. In some cases these are direct analogues of TEI elements, but in other cases they are simplified pseudo-TEI elements that are converted to their more complex equivalents, sometimes with a



1-1 is that of the recapitulatory examination
1-2 otherwise called Quasi Trial.
1-3 Art. 4. The original examination performed
1-4 by the Judge alone: the recapitulatory is per-
1-5 formed by the reconsideration of all evidence

Figure 1: Adding XML-style annotations to Transkribus text.

loss of explicit information. For example, the user might find the <abbrev> element useful for marking up a line of text.

In Figure 1 the transcriber has provided two tags. Line 1–2 contains a technical term tagged with the <term> element, and line 1–3 contains an abbreviated word "Art." (for "Article") tagged as an abbreviation, with the entire word noted using the "expansion" property. The tagging interface would lead the XML-aware user to expect output such as

```
<lb n='2'/>otherwise called <term rend='underline'>Quasi Trial</term>.
<lb n='3'/><abbrev expansion='Article'>Art.</abbrev> 4. The original examination
     performed
```

This, however, is not actually how abbreviations are expressed according to the TEI guidelines. In order to bridge the gap, Transkribus uses an internal XSLT stylesheet to convert the PAGE XML-based transcription data into valid TEI. The conversion results in this:

```
<lb facs='#facs_1_r1l2' n='N002'/>otherwise called <term rend='underline'>Quasi
     Trial</term>.
<lb facs='#facs_1_r1l3' n='N003'/><choice>
<expan>Article</expan>
<abbr>Art.</abbr>
</choice> 4. The original examination performed
```

Now that the abbreviation is represented as a choice, the implicit distinction between the abbreviation (which actually appears in the text) and the expansion (which does not appear as such in the text, but is an interpretation provided by the transcriber by means of a property on the element) has been lost. The distinction, and the information about which of the variants in the `<choice>` element actually appeared in the document, can be arrived at only by understanding the process by which `<abbrev>` became `<choice>`.

The purpose of this seemingly arbitrary dive into the details of how TEI-XML is employed in one particular transcription tool must not be taken as a critique of the software, nor of the decisions the developers made. It is, rather, meant to illustrate a point that is already evident from the recommendations of the TEI Guidelines that a custom schema be produced for each edition project. The point is this: the process of converting text into data does not have a single natural outcome. As such, the technical details of that conversion within any particular software tool cannot be entirely hidden from the user, since these details cannot be inferred by the developers of other software tools that the user might wish to employ. It is the user of the respective tools who must translate the assumptions made by one tool into the assumptions made by the other tool, and this translation can only be done by a user who understands the model and expression of the data she is providing to the tools.

In the case of transcription of manuscript sources, editors who begin by reading the TEI Guidelines will usually adopt an approach that centers around the content of the text and its logical divisions, and treat the visual layout of the text as a secondary concern. On the other hand, editors who wish to have computational assistance with the transcription process—which after all involves the conversion of visual representations of text into records of its logical content—will almost certainly have to work with a form of data that begins with the visual layout, and fits the textual information into that frame. In order for the editor to proceed with further edition of the text, she will need to understand not only what kind of data has been produced, but also the implications of the choices that were made in its production.

4 Text-analysis tools and more complex editions

Eventually the editor will have one or more transcriptions of texts, probably in TEI-XML format. The next step in the process depends very much on the kind of edition being made. For some editions, once the text (or the corpus of texts) is transcribed with appropriate encoding for any notable features, and appropriate commentary or translation encoded alongside in the same document, all that remains is to publish it online in a suitably helpful format.

Many editions, however, are more complex. Some texts may require more structured information, for example to link names and places to relevant external sources of information. Other texts may benefit from algorithmic enrichment, such as morphological analysis, named-entity recognition, or various methods of stylometric analysis. In the case where the same text is preserved in multiple manuscripts, the editor may require a full word-by-word collation of the variant texts and some form of stemmatic analysis.

The multiplicity of different kinds of possible editorial work, and the sheer variety of tools that exist in one form or another to assist with different sorts of work, is one of the great strengths of digital methods for scholarly edition. At the same time, this is by far the greatest bar to existence for the all-inclusive digital edition software package that the scholar may have envisioned. While some of these tools do advertise varying levels of support for TEI-XML encoded texts, others expect to operate only on plain text. The output of these different tools can come in a variety of different forms; the scholar must then decide whether the tool's output should be incorporated somehow into the existing TEI document, or whether the informational content of the edition will have to be stored in multiple complex forms. We can illustrate this with a few examples.

Recogito (Simon et al. 2017) is an online tool for annotating named entities (places, people and events) that appear in a text. Its central feature is the ability to link the place names to entries in online gazetteers such as *Pleiades* (Bagnall et al. 2016). The user may upload texts either as plain text or TEI-XML; in the latter case, there will be a set of assumptions, immutable and invisible to the user, about how to translate the structure defined in a TEI file into 'the text' that should be annotated. Here the scholar will have to exercise caution in ensuring that *Recogito* has, in fact, correctly derived the running text. Once the annotations have been added, the user has a variety of options for extracting them from *Recogito*'s system; these include CSV for tabular data (spreadsheet), RDF for Linked Open Data exchange, formats common in geography and cartography, formats for use as training data in machine learning applications, and, for textual scholars, TEI-XML. The TEI export has certain limitations, however: only place names are exported, and any annotations that would result in overlapping hierarchies (which would cause an XML parser to fail) are not included. The user who would like to enrich a TEI-encoded text with a set of potentially-overlapping semantic annotations prepared in Recogito about all three of people, places, and events, will first need to export these annotations into a separate file in RDF format. Integrating these annotations into the existing TEI document would then be a fairly complex programming task.

Other programs for text analysis, such as the web-based *Voyant* tools (Sinclair, Rockwell, and Voyant Tools Team 2012-) for exploration of corpora with distant-reading techniques, command line tools such as *Mallet* for topic modelling (McCallum 2002), or programming libraries such as the Stylo R package for stylometric analysis (Eder et al. 2016) expect to do their work with plain text supplied by the user. In each of these cases it is up to the user to extract 'the text' to be analyzed from the XML documents that contain them. *Voyant* provides a straightforward and flexible means of doing this directly in their interface, by allowing the user to specify an XPath expression that will extract the text. This, of course, implies that the user is familiar not only with the structure of the XML document in question, but also the XPath syntax for referring to specified portions of the document. Stylo allows import of TEI-XML documents as well as HTML ones, but claims that these input options "have not been extensively tested so far" (Eder et al. 2017). Mallet expects plain text files to be prepared before the tool is run. For any sort of text analysis to have scholarly value, it is clear that the scholar must be able to control the text that is analyzed, ensuring, for example, that transcribers' annotations are not inadvertently included as "original text", or that scribal corrections are handled appropriately. Extraction of the intended text from a TEI-XML file thus involves a set of scholarly decisions that cannot easily be delegated to a non-scholar programmer; as such the scholar must be able to "speak the language" of the model she is using, be it XPath or R.

As a final example, we can take text collation. For editions that involve the reconstruction of a text from multiple manuscript witnesses, there are a number of tools available for collation and comparison of the source texts. Perhaps the two most well-known are *CollateX* (Dekker and Middell 2011), a program that is usually run either from the command line or from within a Python programming environment, and *JuXta* (Performant Software n.d.), a software package with a more user-friendly graphical interface for visualization of variation among texts.

To use *CollateX* it is very important to understand its core data model, which is entirely different from the strictly tree-hierarchical models developed primarily for single-document texts and used by the TEI. Rather, *CollateX* is based on the concept of the *variant graph*, a data structure intended specifically to represent agreement and variation among copies of the same text. The nodes that make up the variant graph—that is, the individual readings in the text—are known as *tokens*, and *CollateX* provides a great deal of flexibility both for the kind of data these tokens can include, and for the criteria that determing whether two tokens ought to be collated with each other. Both the graph and the token structure inform the data output options of *CollateX*. These include tabular formats expressed in CSV and JSON, the graph itself expressed in GraphML, which is an XML format that has no relation to TEI, and a custom form of XML that is modelled on, but does not entirely conform to, the TEI parallel-segmentation model of encoding text variants. Each of these output methods carries with it some trade-off; for example, GraphML is the only format that indicates where *CollateX* has detected transposed readings, but JSON is the only format that preserves extra data that was passed in along with the text of the reading. A textual scholar using *CollateX* must be prepared for these tradeoffs and choose the output format that best matches her needs for further analysis and publication, which itself implies that the scholar must understand what those needs are.

While the core feature of *JuXta* is its visualization options, it also provides a form of output data about the text variation that can be used in further digital processing.

Here the output is a list of variants given in HTML format. Just as with *CollateX*, only the textual content itself is given in the output list; any XML markup in the original TEI document is not preserved. Here too the scholar must evaluate for herself how to use this output, or alternatively, whether to confine the use of *JuXta* to the insights revealed by its visualizations.

The examples given here serve to demonstrate two points. First, the editor has an almost infinite variety of options concerning what to do with the texts that she has transcribed. Just as there is no single set of rules for how to engage in textual criticism, there is no single set of tools that all digital textual scholars will be expected to employ. Second, and following from the sheer variety of methods and needs, the use of any of these tools can only proceed from a solid technical understanding both of the data model that the tool employs, and the interface language needed to communicate with it.

5 From data model to publication

Having read through the section above on text analysis, some editors may at this point be tempted to breathe a sigh of relief. These are the editors whose text is a single document or a collection of similar documents that were transcribed in the same way and using the same model, who have included editorial comments directly in the respective XML document, and who now require only a place to store the TEI file(s) and a suitable means of conversion to a format understood by a modern Web browser.

Throughout Europe there are a growing number of services for storage and preservation of scholarly data, including transcribed text. The two primary options for digital textual scholars based in Austria are the ARCHE repository of the ACDH-ÖAW, and the GAMS repository of the ZIM-ACDH at the University of Graz. Other possibilities include intra-university IT services and commercial hosting, but these do not carry the long-term sustainability benefits that are provided by the dedicated research data repositories.

There are a great many options for conversion of TEI-XML files to a Web-readable format. Perhaps the two simplest, in that they do not require any additional software to run, are the XSLT files provided by the *TEI Boilerplate* project (Walsh, Simpson, and Moaddeli 2012–) and the CSS and Javascript library provided by the *CETEIcean* project (Cayless and Viglianti 2018). Other options, such as *TEIPublisher* (eXist Solutions 2015–), *EVT* (Rosselli Del Turco et al. 2014), and *dsebaseapp* (Andorfer 2016–) are fuller-featured software packages that require the installation and maintenance of software on a Web server. Like the simpler options, these software packages provide reasonable default presentations but, in order to produce an edition that is satisfactory to the scholar, these defaults and the models that inform them will need to be understood and modified.

Although all of these options can be configured easily enough by a technologist who grasps the data model used for text encoding, a word of caution here to the optimistic "non-technical" scholar is in order. Although it is commonplace far beyond the world of digital philology to outsource the development of Web publications to professionals, the user interface to a digital edition communicates a great deal—perhaps more than the unwary scholar realizes—about the context, mood, and import of the text (Andrews and van Zundert 2018; Dillen 2018). The implication here is that, even if the scholar is fortunate enough to have the resources to procure a professional online publication, she will need to understand enough about the possibilities of the medium and about the correspondence between encoded text features and interface functionality to be able to retain control over the message that the interface will broadcast about the text. As such, the scholar who wishes to retain control over the message that is sent by her digital scholarly edition will need to understand the logical models and technical capabilities of the medium in which she publishes.

6 Conclusion

Having made the promised journey through the landscape of tools, we hope not to have lost our readers along the way. The digital medium is extraordinarily full of promise and possibility, and the very possibility that makes digital editions so exciting—the dizzying variety of options and capabilities that could not have been imagined in a print paradigm—also demands a substantial investment of time and understanding. The digital medium is dynamic and interactive in a way that has no parallels in print. It follows that the print paradigm, in which we editors are responsible only for the content which is delivered to a publishing house to take its form, cannot entirely hold. Nevertheless, we continue to hope and trust that the dynamism of the digital will continue to inspire textual scholars to be bold enough to engage with it.

Bibliography

Andorfer, Peter. 2016–. "Dsebaseapp." Accessed September 30, 2019. https://github.com/KONDE-AT/dsebaseapp.

Andrews, Tara L. n.d. "Guidelines for XML File Uploads to Stemmaweb." Accessed September 30, 2019. https://stemmaweb.net/stemmaweb/help/input#cte.

Andrews, Tara L., and Joris J. van Zundert. 2018. "What Are You Trying to Say? The Interface as an Integral Element of Argument." In *Digital Scholarly Editions as Interfaces*, ed. by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider, 3–33. Norderstedt: BoD.

Bagnall, Roger, Richard Talbert, Tom Elliott, Lindsay Holman, Jeffrey Becker, Sarah Bond, Sean Gillies et al. 2016–. *Pleiades: A Gazetteer of Past Places.* Accessed September 30, 2019. https://pleiades.stoa.org.

Bordalejo, Barbara. 2018. "Digital versus Analogue Textual Scholarship or The Revolution Is Just in the Title." *Digital Philology: A Journal of Medieval Cultures* 7: 7–28. doi:10.1353/dph.2018.0001.

Causer, Tim, and Valerie Wallace. 2012. "Building A Volunteer Community: Results and Findings from Transcribe Bentham." *Digital Humanities Quarterly* 6 (2). Accessed October 1, 2019. http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html.

Cayless, Hugh, and Raffaele Viglianti. 2018. "CETEIcean: TEI in the Browser." In *Proceedings of Balisage: The Markup Conference 2018*, Balisage Series on Markup Technologies, vol. 21, Washington, DC. doi:10.4242/BalisageVol21.Cayless01.

Dekker, Ronald Haentjens, and Gregor Middell. 2011. "Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements." Paper presented at the meeting *Supporting Digital Humanities 2011.* Copenhagen.

Dillen, Wout. 2018. "The Editor in the Interface: Guiding the User through Texts and Images." In *Digital Scholarly Editions as Interfaces*, ed. by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider, 35–59. Norderstedt: BoD.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1): 107–21.

Eder, Maciej, Mike Kestemont, Jan Rybicki, and Steffen Pielström. 2017. "'Stylo': R Package for Stylometric Analyses." Accessed September 30, 2019. https://github.com/computationalstylistics/stylo.

eXist Solutions. 2015–. "TEI Publisher." Accessed September 30, 2019. https://teipublisher.com.

Haentjens Dekker, Roland, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom, and David J. Birnbaum. 2018. "TAGML: A Markup Language of Many Dimensions." In *Proceedings of Balisage: The Markup Conference 2018*, Balisage Series on Markup Technologies, vol. 21, Washington, DC. doi:10.4242/BalisageVol21.HaentjensDekker01.

Hagel, Stefan. 1997–. "Classical Text Editor.' Accessed September 30, 2019. https://cte.oeaw.ac.at.

—. 2007. "The Classical Text Editor. An Attempt to Provide for Both Printed and Digital Editions." In *Digital Philology and Medieval Texts*, ed. by Arianna Ciula, and Francesco Stella, 77–84. Accessed Sept. 30, 2019. http://www.infotext.unisi.it/upload/DIGIMED06/book/hagel.pdf.

"How to Use Transkribus—in 10 Steps (or Less)." 2015. Accessed September 30, 2019. https://transkribus.eu/Transkribus/docs/How%20to%20use%20TRANSKRIBUS-0.1.5.pdf.

Institut für neutestamentliche Textforschung. n.d. "New Testament Virtual Manuscript Room — INTF." Accessed October 1, 2019. http://ntvmr.uni-muenster.de/home.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. "Transkribus — A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, 19–24. doi:10.1109/ICDAR.2017.307.

Klug, Helmut W., Selina Galka, and Elisabeth Steiner, eds. 2021. "Tools." In *Weißbuch Digitale Editionen.* Graz: HRSM-Projekt Kompetenznetzwerk Digitale Editionen. https://www.digitale-edition.at/archive/objects/context:konde/methods/sdef:Context/get?mode=tools.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." Accessed September 30, 2019. http://mallet.cs.umass.edu.

Mittelbach, Arno, Sebastian Rahtz, and Ioan Bernevig. 2018. "Roma: Generating Customizations for the TEI (version 5.0.0)." Accessed September 30, 2019. https://roma2.tei-c.org.

Performant Software. n.d. "Juxta: Collation Software for Scholars." Accessed September 30, 2019. https://web.archive.org/web/20220307201635/http://juxtasoftware.org.

Piez, Wendell. 2014. "TEI in LMNL: Implications for Modeling." *Journal of the Text Encoding Initiative* 8 (December). doi:10.4000/jtei.1337.

Pletschacher, Sefan, and Apostolos Antonacopoulos. 2010. "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework." In *2010 20th International Conference on Pattern Recognition*, 257–60. doi:10.1109/ICPR.2010.72.

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 (December). doi:10.4000/jtei.1077.

Sahle, Patrick. 2016. "What Is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, ed. by Matthew James Driscoll, and Elena Pierazzo, 19–40. Open Book Publishers. doi:10.11647/OBP.0095.02.

Simon, Rainer, Elton Barker, Leif Isaksen, Pau De Soto CaÑamares. 2017. "Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2." *Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives* 13: 111–32. doi:10.1080/15420353.2017.1307303.

Sinclair, Stefan, Geoffrey Rockwell, and Voyant Tools Team. 2012–. "Voyant Tools." Accessed October 1, 2019. https://voyant-tools.org.

Vogeler, Georg. 2019. "Digitale Editionspraxis. Vom pluralistischen Textbegriff zur pluralistischen Softwarelösung." In *Textgenese in der digitalen Edition*, ed. by Anke Bosse, and Walter Fanta, 117–36. Berlin, Boston: De Gruyter. doi:10.1515/9783110575996-008.

Walsh, John, Grant Simpson, and Saeed Moaddeli. 2012–. "TEI Boilerplate." Accessed October 1, 2019. http://teiboilerplate.org.

Zundert, Joris van, and Peter Boot. 2011. "The Digital Edition 2.0 and the Digital Library: Services, Not Resources." In *Digitale Edition Und Forschungsbibliothek*, ed. by Christiane Fritze, Franz Fischer, Patrick Sahle, and Malte Rehbein, 141–52. Wiesbaden: Harrassowitz.