UNIVERSITÄT GRAZ
UNIVERSITY OF GRAZ

What is text, really? TEI and beyond

September 16 - 20, University of Graz, Austria

Book of Abstracts

CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

What is text, really? TEI and beyond

Preface to the Book of Abstracts

This PDF contains the abstracts for the 19th annual Conference and Members' Meeting of the Text Encoding Initiative Consortium (TEI), organized by the Centre for Information Modelling, University of Graz, Austria. September 16–20, 2019. It is a rough compilation of the PDF/word-processor files we received by the authors. Not every text file contains authors, so please check the authors as given in the table of content in the following pages. We consider the XML/TEI archived at https://gams.uni-graz.at/context:tei2019 as the main reference to each contribution.

The 2019 edition of the annual TEI conference is put under the theme "What is text, really? TEI and beyond". The development of the TEI and the discussions around it have shown how manifold text(s) can be, and how far-reaching the TEI approach is. Thus, this year's theme "What is text, really?" poses a fundamental question, which goes beyond the pure reference to the seminal paper by DeRose, Durand, Mylonas, and Renear. In 1990 they answered the question somewhat pragmatically introducing a model for text as an ordered hierarchy of content objects which can easily be formalised with digital technologies, but, as they said later on: text can be much more than that. Text encoding can make various aspects of texts explicit, enabling scholars to examine their nature and their relationship with other objects. In this context, the power of the TEI relies on its technological interactions, supporting software of all kinds operating upon texts, from visualisation to annotation tools, digital publishing systems, or statistical analysis. The TEI framework is a way of modelling knowledge and engaging in a dialogue with ontologies, conceptual models, and recent approaches such as text as graph. For text-centric disciplines the TEI offers a range of solutions that address core research needs. However, for object-based disciplines, like archaeology or museology, where text and its encoding is only a small part of their data modelling ecosystem, the value of TEI is not so clear and it competes with other modeling approaches.

This digital resource is intended to create a track of the intellectual input of the event in the digital realm. The TEI documents handed-in by the authors did not follow any other rule than to be TEI compliant. We did not receive customized documents, but converted 34 from standard word processing formats into TEI thanks to Oxgarage. The outcome is as heterogenuous as the application and interpretation of the TEI guidelines are. We normlized title, author, and keywords from the conftool submissions. This digital publication only starts the path to your own work with the corpus: you are invited to extract keywords, author affiliations, or bibliography, analyze style or tag usage in the very texts at your own wish.

All abstracts collected here were handed-in as 300 word proposals for evaluation by the at least three reviewers from program committee. The authors took account of the review results in final versions of their contributions and could extend their abstract. Here you find this final submission.

We want to thank the sponsors (SyncroSoft, the Dean of the Faculty for Humanities, the University of Graz, the City of Graz, CLARIAH.AT) for their financial support. We thank the TEI community for their creative response to the call and their active participation. Without the collective widsom of the program committee, this book of abstracts would have much less quality: thanks Susanna, Syd, Elisa, Vanessa, Roman, Hugh, James, Frank, Franz, Julia, Christiane, Susanne, Martin, Luis, Kiyonori, Claudia, Laurent, Walter, Martina, José, Kathryn, Magdalena, Pip, and in particular my co-chair Gimena!! However, in this place the deepest gratititude goes to those who converted this book of abstract from an idea into reality: Elisabeth Raunig, Elisabeth Steiner, Martina Scholger for management, copy editing, and digital publication, and for the content - of course - the authors.

Graz, 2019, September 12th

Georg Vogeler

Table of Content

Papers:

120, Cummings, James, Highlighting Our Examples: encoding XML examples in pedagogical contexts

121, Cummings, James, Introducing Objectification: when is an <object> a <place>?

122, Haaf, Susanne, Exploring TEI structures to find distinctive features of text types

123, Dumont, Stefan; Grabsch, Sascha; Müller-Laackman, Jonas, correspSearch v2 – New ways of exploring correspondence

124, Pytlik Zillig, Brian L; Bolin, Mary, The Semantic Field Grace in Early Modern English

125, Creamer, Andrew; Lembi, Gaia; Mylonas, Elli; Satlow, Michael, Archiving a TEI project FAIRly

127, Bauman, Syd, Validating @selector: a regular expression adventure

128, Dumont, Stefan; Haaf, Susanne; Seifert, Sabine, TEI encoding of correspondence: A community effort

129, Janssen, Maarten, Advantages and challenges of tokenized TEI

131, Gibson, Nathan P., Five Centuries of History in a Network

134, Prager, Katharina; Hannesschlaeger, Vanessa; Boerner, Ingo, Karl Kraus contra …, or: text contra action

136, Mörth, Karlheinz; Schopper, Daniel, Modelling linguistic knowledge in TEI: the case of the Vienna Corpus of Arabic Varieties

138, Roeder, Torsten, Genesis and Variance: From Letter to Literature

139, Diehr, Franziska ; Gronemeyer, Sven; Sikora, Uwe; Prager, Christian; Brodhun, Maximilian; Wagner, Elisabeth; Diederichs, Katja; Grube, Nikolai, Inscriptions, Hieroglyphs, Linguistics… and Beyond! The Corpus of Classic Mayan as an Ontological Information Resource

140, Šimek, Jakub, Referencing an editorial ontology from the TEI: An attempt to overcome informal typologies

141, Fritze, Christiane; Klug, Helmut; Kurz, Stephan; Steindl, Christoph, Recreating history through events

145, Cole, Nicholas; De Roure, David; Willcox, Pip, TEI XML and Delta Format Interchangeability

146, Gfrörer, Samuel; Thoden, Klaus, Case Study TEI Customization: A Restricted TEI Format for Edition Open Access (EOA)

148, Granholm, Patrik; Olsson, Leif-Jöran, Manuscripta - The editor from past to future

151, Bowers, Jack; Stöckle, Philipp; Breuer, Hans Christian; Breuer, Ludwig Maximilian, Native-TEI dialectal dictionary for Bavarian dialects in Austria: data structure, software and workflow

155, Okada, Kazuhiro; Nakamura, Satoru; Nagasaki, Kiyonori, An Encoding Strategic Proposal of “Ruby” Texts: Examples from Japanese Texts

156, Nakamura, Satoru; Okada, Kazuhiro; Nagasaki, Kiyonori, An Attempt of Dissemination of TEI in a TEI-underdeveloped country: Activities of the SIG EAJ

157, Erjavec, Tomaž; Pančur, Andrej, Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings

158, Mondaca, Francisco; Schildkamp, Philip; Rau, Felix; Bigalke, Jan, Introducing an Open, Dynamic and Efficient Lexical Data Access for TEI-encoded Dictionaries on the Internet

159, Pollin, Christopher; Tomasek, Kathryn, Making Linkable Data from Account Books: Bookkeeping Ontology in the Digital Edition Publishing Cooperative for Historical Accounts

160, Maximova, Daria; Fischer, Frank, Using Machine Learning for the Automated Classification of Stage Directions in TEI-Encoded Drama Corpora

164, Cugliana, Elisa; Barabucci, Gioele, A sign of the times: medieval punctuation, its encoding and its rendition in modern times

165, Wissik, Tanja, Challenges in encoding parliamentary data: between applause and interjections

166, Lehečka, Boris, Using Microsoft Word for preparing XML TEI-compliant digital editions

168, Benito, Alejandro; Doran, Michelle; Edmond, Jennifer; Kozak, Michał; Mazurek, Cezary; Rodríguez, Alejandro; Therón, Roberto; Wandl-Vogt, Eveline, Analyzing and Visualizing Uncertain Knowledge: Introducing the PROVIDEDH Open Science Platform

169, Elagina, Daria, Case Study: Ethiopian Psalter of the Virgin

170, Josfeld, Julia; Simpson, Grant Leyton, Reflecting the Influence of Technology on Models of Text in Scholarly Digital Editing

171, Daengeli, Peter; Forney, Christian, Referencing annotations as a core concept of the hallerNet edition and research platform

172, Hinkelmanns, Peter, Text Graph Ontology. A Semantic Web approach to represent genetic scholarly editions

173, Roberts-Smith, Jennifer; Takeda, Joseph; Kaethler, Mark; Malone, Toby; Jenstad, Janelle, Reconceiving TEI models of theatrical performance text with reference to promptbooks

174, Takeda, Joseph; Lines, Sydney, Using Github and its Integrations to Create, Test, and Deploy a Digital Edition

176, Salgado, Ana; Costa, Rute; Tasovac, Toma, TEI Lex-0: a good fit for the encoding of the Portuguese Academy Dictionary?

177, Mueller, Ermenegilda Rachel, A TEI customization for the description of paper and watermarks

178, Khemakhem, Mohamed; Galleron, Ioana; Williams, Geoffrey; Romary, Laurent; Suárez, Pedro Ortiz, How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures

180, Jenstad, Janelle Auriol; Takeda, Joseph; Greatley-Hirsch, Brett; Mardock, James, What is a Line? Encoding and Counting Lines in Early Modern Dramatic Texts

187, McAllister, Patrick, Growing collections of TEI texts: Some lessons from SARIT

188, Trikha, Himal, Towards larger corpora of Indic texts: For now, minimize metatext

189, Fermer, Mathias, Encoding history in TEI: A corpus-oriented approach for investigating Tibetan historiography

Panels:

132, Bleier, Roman; Fischer, Franz; Gengnagel, Tessa; Klug, Helmut W.; Sahle, Patrick; Steiner, Christian; Winslow, Sean M.; Worm, Andrea, Graphs - charters - recipes: challenges of modelling medieval genres with the TEI

144, Jones, Huw Eifion; Faghihi, Yasmin; Holford, Matthew; Freeman, James; Schassan, Torsten; Guariento, Luca, TEI for manuscript description: progress, problems, and potential

150, Beshero-Bondar, Elisa; Cayless, Hugh; Viglianti, Raffaele; Cummings, James, Document Modeling with the TEI Critical Apparatus

Posters:

105, Albarrán-Fernández, Elena, A TEI-based model to encode notarial charters (Asturias, 1260-1350 ca.)

113, Raybuck, Suzanne Michelle; Stanley, Sarah C., An Exploration of <object> Using Antarctic Artifacts

118, Del Grosso, Angelo Mario; Spampinato, Daria; Capizzi, Erica; Cristofaro, Salvatore; Seminara, Graziella, Promoting Bellini’s legacy and the Italian opera by scholarly digital editing his own correspondence

130, Steiner, Christian; Klug, Helmut W.; Böhm, Astrid; Raunig, Elisabeth ; Laurioux, Bruno; Ardesi, Denise; Poirier, Corentin, Annotating Cooking Recipes of the Middle Ages for semantic analysis and visualization

137, Bleier, Roman; Brantner, Elisabeth; Haug-Moritz, Gabriele; Neerfeld, Christiane; Ortlieb, Eva; Schreiber, Thomas; Vogeler, Georg; Zeilinger, Florian, Encoding the documents of the “The Imperial Diet in Regensburg, 1576”

147, Modráková, Renáta; Paličková, Tereza, Virtual reconstruction of scattered provenance of Bohemian printed books

149, Modráková, Renáta, Research of provenance glosses in medieval manuscripts and in incunabulas from historical collections of the National Library of the Czech Republic

153, Alles Torrent, Susanna; del Rio Riande, Gimena, TTHUB: Text Technologies Hub for Extending TEI Training in Spanish

185, Zimmer, Mary Erica; Jenstad, Janelle Auriol, Documenting Discoveries: TEI and Browsing the Bookshops in Paul’s Cross Churchyard

186, Fuchs, Alexandra; Geiger, Bernhard; Hobisch, Elisabeth; Koncar, Philipp; Scholger, Martina; Saric, Sanja , Distant Spectators: Mining TEI-encoded periodicals of the Enlightenment

193, Viglianti, Raffaele, A new Roma (beta): a rich interface for ODD customization

194, Busch, Anna; Fischer, Frank, How to encode the unsaid with the TEI

Workshops:

133, Maus, David, An introduction to Schematron and Schematron QuickFix

162, Viglianti, Raffaele; Cayless, Hugh, Minimalist TEI Publishing with CETEIcean (/sɪˈtiːʃn/)

179, Kampkaspar, Dario, Interface design and user involvement – Wiennerisches Diarium Digital

181, Turska, Magdalena; Meier, Wolfgang, Hands-on TEI publishing

Demonstrations:

112, Pellissier Tanon, Thomas, TEI MediaWiki extension

126, Kuczera, Andreas, TEI as a Graph

142, Janssen, Maarten, TEITOK – TEI based annotated corpora

143, Bauman, Syd, Character counting

175, Gimena, Maria; Simon, Rainer; Barker, Elton; Isaksen, Leif; Kahn, Rebecca; Vitale, Valeria; Rojas Castro, Antonio; Cayless, Hugh, Recogito: from Semantic Annotation to Digital Scholarly Edition

183, Robinson, Peter, Textual Communities

184, Turska, Magdalena; Meier, Wolfgang, Van Gogh Letters: the TEI Publisher clone

Papers

A realistic theory of textuality and its consequences on digital text representation

Author: Fabio Ciotti

The theoretical debate about digital textuality in the last decades has been deeply influenced by the post-modernist theory. If this influence was quite apparent in the debates about the hypertext, we can find some of its fundamental tenets also in text encoding and digital scholarly editing theories (Landow 1997; McGann 2001). As a consequence, even in these areas we can find a general and strong support to anti-realist or constructivist notions about textuality and its digital representation (Patrick Shale 2013a and 2013b; Ciula and Marras 2016). These views oscillate from radical ontological stances, epitomized by this notorious McGann's sentence: 'What is text? I am not so naive as to imagine that question could ever be finally settled. Asking such a question is like asking 'How long is the coast of England?'' (McGann 2002); to weaker epistemological or pragmatist stances that advocate for the plurality of the textual re-representations (the double *re* is due to the fact that a text is in itself a representation).

In this paper I propose a weak realist theory of (digital) textuality (somehow along the lines or the *new realism* movement in the recent philosophical debate (Ferraris, Bilgrami, and De Caro 2012)) building on the theory of notation developed by Nelson Goodman in its *Languages of Art* (Goodman 1968) and the theory of intentional systems devised by Daniel Dennet (Dennet 1987 and for an introductory recapitulation Dennett 2009). In brief: texts are spatiotemporal artifacts that have causal roles in our cognitive understanding. As Dennet explains, we as a species, tend to adopt the *intentional stance* to explain the rationales and functions (or meaning) of complex systems, that consists in attributing meaning and beliefs to those artifacts. Documents are *prima facie* interpreted as intentional artifacts that convey meanings. At this level what really counts is the possibility to identify and fix the notational nature of the text, because it's primarily that notational nature that has a causal role in the chain that starts form perception and ends in the cognitive work of interpretation (meaning attribution).

The general nature of the notational systems has been analyzed by another philosopher of the preceding generation, Nelson Goodman. In his influent book (for many and diverse reasons) *The languages of Art, he affirms* that textual artifacts have the distinctive property to be 'in a definite notation, consisting of certain signs or characters that are to be combined by concatenation' (Goodman 1968, 116); this *notational condition* provides a *principium individuationis* for the text.

Hence, a (digital) representation of a text is adequate if and only if it has the property of having the 'sameness of spelling': the exact correspondence as sequences of letters, spaces, punctuation marks. All the other properties are either contingent (if they are material) or derivative (all the cognitive objects produced in the act of reading). Of course, one can or will try to adopt the other two level of explanation identified by Dennett (the *design stance* and the *physical stance*, that can be translated in our context into various technical approaches in textual studies), but this happens only for limited occasions and in particular conditions.

Moving from this general theoretical account most of the recurring debates about the pros and cons of one encoding metalanguage or digital modeling strategy over another can be reframed into a moderate pluralistic methodological framework, where the unique central point is the correct (as far as possible) representation of the characters sequences of the textual artifacts that are generally used as text-documents of a specific text-work, to which all other properties can be attached. I tend to conceive a full-fledged stand-off markup strategy as the most natural way of giving a computational modeling of the textual reality, but other strategies and notational approaches (like XML inline markup) can have technical, pragmatic and cultural affordances that, under many respects, are advantageous.

Getting Along with Relational Databases

Martin Holmes

ABSTRACT

Both relational databases and XML have strengths and weaknesses as data storage and modelling systems. Most researchers working with Humanities historical and literary data would argue for the superiority of XML, since it allows unlimited nesting, linking, and complexity. RDB proponents claim superior querying and processing speed, although recent advances in XML languages and tools have eroded that advantage.

Nevertheless, RDBs remain popular, and many researchers seem instinctively to prefer them. Most DH programmers have encountered researchers who know little about databases or data modelling, but are nevertheless convinced that what they need and must have for their project is a database. Databases are somehow compelling and attractive in a way that XML is not. Perhaps the familiarity of tabular data representations is comforting; maybe forcing data into constrained representations seems to constitute mastering it somehow.

So, sometimes against our better judgement or advice, a project may end up with both an RDB and an XML document collection, and programmers must then integrate these distinct forms of data when building project outputs. This presentation discusses the Digital Victorian Periodical

Poetry (DVPP) project, where metadata about 15,000 poems from nineteenth-century periodicals is captured in a MySQL database, and periodically exported to create a TEI file for each poem. Many of the poems are then transcribed and encoded. The canonical source of metadata is the RDB, while the canonical source of textual data is the TEI file. Metadata in the TEI files must be periodically updated from the RDB, without disturbing the textual encoding. Changes to the RDB data may result in changes to the id and filename of the related TEI file, so any existing TEI data is migrated to a new file, and the SVN repository must be appropriately updated. All of this is done with XSLT and Ant.

INDEX

Keywords: TEI and non-XML technologies, TEI and beyond: interactions, interchange, integrations and interoperability, TEI environments and infrastructures

1. Background

1 Relational databases (RDBs) and XML are both mature technologies that have been in common use for decades. It is arguable that they arise out of the same roots. Early work on data storage and modelling in the 1960s gave rise to IBM's mainframe database management system IMS, which represented data in the form of hierarchical trees. C.J. Date's (1991) classic *An Introduction to Database Systems* has an Appendix devoted to IMS which describes it in terminology that would be familiar to any XML encoder. IMS even addresses the perennial issue of overlapping hierarchies, by allowing "A secondary data structure" which is "still a hierarchy, but a hierarchy in which participant segments have been rearranged, possibly drastically"; in other words, it allows for multiple hierarchies over the same dataset. However, beginning with the work of E.F. Codd in the 1970s and the rise of SQL, the relational database model familiar today became dominant, and remained so until the relatively recent popularity of NoSQL approaches.

2 Both relational databases and XML have strengths and weaknesses as data storage and modelling systems. Most researchers working with Humanities historical and literary data would argue for the superiority of XML, since it allows unlimited nesting, linking, and complexity. RDB proponents claim superior integrity constraints, querying, and processing speed, although recent advances in XML languages, database engines and tools have somewhat eroded those advantages.

3 Nevertheless, RDBs remain popular, and many researchers seem instinctively to prefer them. Most digital humanities programmers have encountered researchers who know little about databases or data modelling, but are nevertheless convinced that what they need and must have for their project is a database. Databases are somehow compelling and attractive in a way that XML is not. Perhaps the familiarity of tabular data representations is comforting; maybe forcing data into constrained representations seems to constitute mastering it somehow; or perhaps the tendency to gather initial data in the early stages of a project using spreadsheets, for want of a better tool, encourages conception of data (especially metadata) in terms of columns and rows. Whatever the reason, in one way or another, sometimes against our better judgement or advice, a project may end up with both an RDB and an XML document collection, and programmers must then integrate these distinct forms of data when building project outputs.

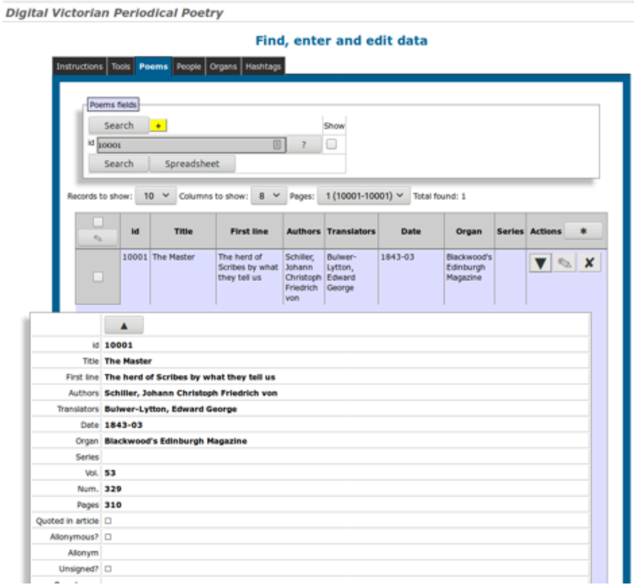
4 Approaches to integrating RDB and XML data have normally taken the form of storing XML data into RDB fields, and then providing some level of richer access to that data through the use of XPath or XQuery (see Bertino and Catania for a useful overview). This is the approach taken by the ReMetCa team (González-Blanco and Rodríguez 2015): XML fragments representing verse (not full documents) are stored in text fields in a relational database, and the relationships between them are modelled using the RDB schema. However, such an approach is far from ideal; González-Blanco and Rodríguez describe some of the limitations and frustrations they encountered in modelling the poetic structure of the verse in their database; they struggled with "a complex model of relationships among those components which are very difficult to represent in a database," and they conclude that "the E-R model is inappropriate for this purpose due to its center-based structure, with the entities of poem, line, and stanza in the middle of its referential domain of study" (para 8). Gibson (2012) describes a similar scenario with mixed RDB and XML data, and how he used Saxon's SQL extension functions to overcome the problem.

5 However, storing XML data in RDB fields is suboptimal. Most serious encoding projects make use of version-control systems such as Git or Subversion, for very good reasons: in a project with many transcribers and encoders, where multiple waves of encoding and annotation may be applied to each document, it is essential to maintain a detailed revision history which makes it possible to recover any previous incarnation of any document, and to track the revisions made to specific parts of the document by specific encoders.

2. The *Digital Victorian Periodical Poetry* project

6 This presentation will focus on the integration of RDB and XML data in the *Digital Victorian Periodical Poetry* project. This project began life many years ago as a pure-metadata project, capturing information about tens of thousands of poems that appeared in British periodicals during the nineteenth century. At that time, an RDB system seemed a natural and sufficient tool for the job, so a MySQL database, along with a data-entry interface, was set up for the researchers, and data collection proceeded rapidly (figure 1). However, after some years the project gained an additional research focus, and, more recently, funding from the Social Sciences and Research Council of Canada, to transcribe and encode a subset of these poems; we are focusing primarily on the decade years (1820, 1830, 1840 and so on through to 1900), and we expect to encode around 2,000 poems. Meanwhile, indexing of the much larger dataset continues.

Figure 1. A record in the relational database.



id 10001 □

Records to show: 10 Columns to show: 8 Pages: 1 (10001-10001) Total found: 1

□	Id	Title	First line	Authors	Translators	Date	Organ	Series	Actions
□	10001	The Master	The herd of Scribes by what they tell us	Schiller, Johann Christoph Friedrich von	Bulwer-Lytton, Edward George	1843-03	Blackwood's Edinburgh Magazine		▼ ✎ ✗

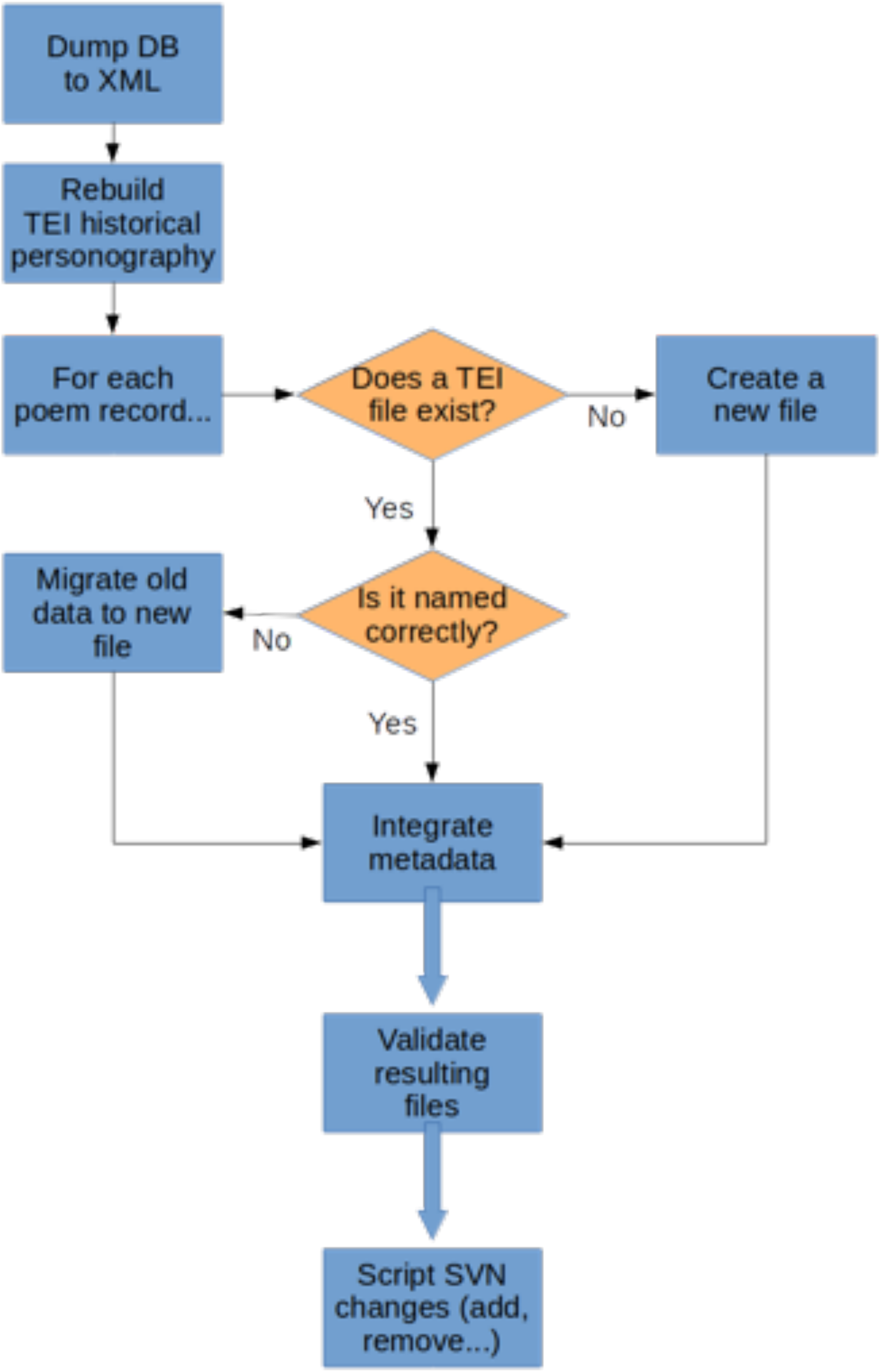
Id	10001
Title	The Master
First line	The herd of Scribes by what they tell us
Authors	Schiller, Johann Christoph Friedrich von
Translators	Bulwer-Lytton, Edward George
Date	1843-03
Organ	Blackwood's Edinburgh Magazine
Series	
Vol.	53
Num.	329
Pages	310
Quoted in article	□
Allonymous?	□
Allonym	
Unsigned?	□

7 Our long-term plan is for the entire dataset to be in the form of TEI XML files, but for the next few months, data will continue to be added to the RDB system, since we have good methods and protocols for this, as well as trained research assistants who are used to working with it. We are now also well into our encoding process, and for that we need to generate individual TEI files for each poem, and store them in Subversion.

8 In this hybrid project, the canonical source of metadata for the poems is the RDB, while the canonical source of textual data is the TEI XML files. To build and test the project outputs, we need to generate TEI files for every poem, whether or not it has, or will have, an encoded transcription. The metadata stored in the TEI files must be periodically updated based on the RDB, without disturbing any of the textual encoding or the additional metadata in the TEI header relating specifically to the encoding (responsibility statements, rendition elements, category references, and so on). Changes to the RDB data may result in changes to the id and filename of the related

TEI file, so any existing TEI data must be migrated to a new file, and the SVN repository must be appropriately updated. The presentation will describe how this process is accomplished safely without loss of data, using a system based primarily on Apache Ant and XSLT (figure 2).

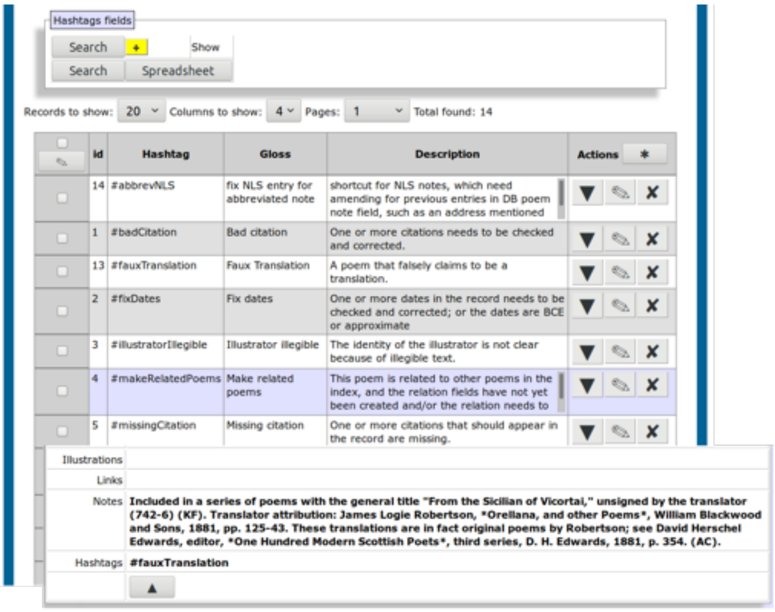
Figure 2. A simple representation of the metadata integration process.



Normally, we run the database integration process only on a small subset of the data at one time; for example, we may refresh the metadata in all the poems from a specific periodical in a specific year, in preparation for the transcription/encoding team starting work on that year.

9 By the end of 2019, we plan to eliminate the relational database entirely. Although it is a convenient tool for collecting metadata while working through large numbers of periodicals, its limitations are constantly frustrating; every day we encounter situations in which something relatively trivial to encode in TEI would require substantial modification to the structure and complexity of the database. For example, degrees of uncertainty about the identity of an author, or about whether two pseudonyms represent the same person, can easily be expressed in TEI, but require additional joining tables in the database. Similarly, some poems claim to be translations but are probably not, and their "translators" are probably their authors. Ambiguities such as this are difficult to handle in a categorical system such as an RDB, but they are the bread and butter of TEI encoding.

Figure 3. Use of the "hashtag" system for more flexible database entries.



Records to show: 20 Columns to show: 4 Pages: 1 Total found: 14

□	id	Hashtag	Gloss	Description	Actions
□	14	#abbrevNLS	fix NLS entry for abbreviated note	shortcut for NLS notes, which need amending for previous entries in DB poem note field, such as an address mentioned	▼ ✎ ✗
□	1	#badCitation	Bad citation	One or more citations needs to be checked and corrected.	▼ ✎ ✗
□	13	#fauxTranslation	Faux Translation	A poem that falsely claims to be a translation.	▼ ✎ ✗
□	2	#fixDates	Fix dates	One or more dates in the record needs to be checked and corrected; or the dates are BCE or approximate	▼ ✎ ✗
□	3	#illustratorIllegible	Illustrator illegible	The identity of the illustrator is not clear because of illegible text.	▼ ✎ ✗
□	4	#makeRelatedPoems	Make related poems	This poem is related to other poems in the index, and the relation fields have not yet been created and/or the relation needs to	▼ ✎ ✗
□	5	#missingCitation	Missing citation	One or more citations that should appear in the record are missing.	▼ ✎ ✗

Illustrations	
Links	
Notes	**Included in a series of poems with the general title "From the Sicilian of Vicortai," unsigned by the translator (742-6) (KF). Translator attribution: James Logie Robertson, *Orellana, and other Poems*, William Blackwood and Sons, 1881, pp. 125-43. These translations are in fact original poems by Robertson; see David Herschel Edwards, editor, *One Hundred Modern Scottish Poets*, third series, D. H. Edwards, 1881, p. 354. (AC).**
Hashtags	**#fauxTranslation**
	▲

10 Meanwhile, we live with the database, and devise new cunning strategies to make it more bearable. The constantly-shifting requirements of the metadata team, as they encounter new features and unexpected exceptions in the incoming data, have led to the development of an ad-hoc system

based on “hashtag” fields, where researchers can use Twitter-style hashtags in free text notes fields to capture things that would otherwise require database extension or modification. Hashtags are defined in a separate table, and new ones can be added at will, but their use is policed by a diagnostic process that identifies any instances of hashtags in the notes fields which are not in the hashtag table, thereby avoiding the proliferation of typographical errors or the use of undefined hashtags. The hashtags are themselves transformed into a taxonomy in the TEI data, and from there they are integrated into the project schema as values for `@target` attributes on `<catRef>` elements. Figure 3 shows the hashtag table, along with an instance of a hashtag in use in a pooem record. The Notes field above it shows more evidence of data that really needs rich encoding for titles, dates, names and so on, but must, for the moment, be handled with plain text.

BIBLIOGRAPHY

Bertino, E., and B. Catania. 2001. “Integrating XML and databases.” *IEEE Internet Computing* (5:4) 84–88. https://doi.org/10.1109/4236.939454.

Date, C.J. 1991. *An Introduction to Database Systems*. Vol. 1. 5th Edition. Reading, Mass.: Addison-Wesley.

Gibson, Matthew. 2012. “Using XSLT’s SQL Extension with Encyclopedia Virginia.” *Code{4}lib Journal* 16. https://journal.code4lib.org/articles/6486.

González-Blanco, Elena, and José Luis Rodríguez. 2015. “ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse.” *Journal of the Text Encoding Initiative*, Issue 8. http://journals.openedition.org/jtei/1274. DOI : 10.4000/jtei.1274.

AUTHOR

MARTIN HOLMES

Programmer/Consultant, University of Victoria Humanities Computing and Media Centre

1 Abstract

Any expansion of the TEI beyond its traditional user-base involves a recognition of differing answers to the traditional question ‘What is text, really?’ [[8]; [7]; [13]], and hence a rethinking of some aspects of TEI praxis. We report on work carried out in the context of the COST Action CA16204 "Distant Reading", in particular on the TEI-conformant schemas developed for one of the Action's principal deliverables: the European Literary Text Collection (ELTeC).

The ELTeC will contain comparable corpora for each of a dozen European languages, each being a balanced sample of 100 novels from the 19th century, together with metadata situating them in their contexts of production and of reception. We hope that it will become a reliable basis for comparative work in data-driven textual analytics, enabling researchers to go beyond a simple ‘bag of words’ approach, while respecting views of "what text really is" currently dominant in such fields as statistically-derived authorship attribution, topic modelling, character network analysis, and stylistic analysis in general.

The focus of the ELTeC encoding scheme is not to represent texts in all their original complexity, nor to duplicate the work of scholarly editors. Instead, we aim to facilitate a richer and better-informed distant reading than a transcription of lexical content alone would permit. Where the TEI permits diversity, we enforce consistency, by defining encodings which permit only a specific and quite small set of textual features, both structural and lexical. We also define a single TEI-conformant way of representing the results of textual analyses such as named entity recognition or morphological parsing, and a specific set of metadata features. These constraints are expressed by a master TEI ODD, from which we derive three different schemas by ODD-chaining, each associated with appropriate documentation.

Lou Burnard is an independent consultant in TEI XML. He was for many years Associate Director of Oxford University Computing Services, and was one of the original editors of the TEI.

Christof Schöch is Professor of Digital Humanities at the University of Trier, Germany, and Co-Director of the Trier Centre for Digital Humanities (TCDH). He chairs the COST Action "Distant Reading for European Literary History" (CA16204).

Carolin Odebrecht is a corpus linguist at Humboldt-Universität zu Berlin. Her research fields are modelling, creating, archiving of historical corpora and corpus metadata.

In search of comity: TEI for distant reading

Lou Burnard, Christof Schöch and Carolin Odebrecht

2 Introduction

Comity is a term from theology or political studies, where it is used to describe the formal recognition by different religions, nation states, or cultures that other such entities have as much right to existence as themselves. In applied linguistics, the term has also been used by such writers as Widdowson or Aston [[2]] seeking to demonstrate how the establishment of comity can facilitate successful inter-cultural communication, even in the absence of linguistic competence[1]. We appropriate the term here in this latter sense, as a means of re-asserting the inter-disciplinary roots of the TEI.

Recent histories of the TEI (e.g. [9]) have a tendency to under-emphasize the multiplicity of disciplines gathered at its birth, preferring to focus on those disciplines which can be plausibly framed as prefiguring our current configuration of the 'digital humanities' in some way. Yet the Poughkeepsie conference and the process of designing the Guidelines which followed alike were kickstarted by input from corpus linguists and computer scientists as well as traditional philologically-minded editors and source-driven historians. The TEI belongs to a multiplicity of research communities, dating as it does from a period when computational linguists and traditional philologists alike were beginning to wake up to the implications of the advent of massive amounts of digital text for their disciplines. The steering committee which oversaw its development and the TEI editors alike conscientiously attempted to ensure that the Guidelines should reflect a view of text which was generally shared and generic, rather than specific to any discipline or to any particular usage model.

The TEI's radical proposition that there was such a thing as a single abstract model of textual components, which might usefully be considered independently of its expression in a particular source or output, or its use in any particular discipline, was necessarily at odds with at least two prevailing orthodoxies: on the one hand, the view that a text is no less and no more than the physical documents which instantiate it, and can be adequately described and represented by its salient visual properties alone; on the other hand, the view

[1] 'Those participating in conversational encounters have to have a care for the preservation of good relations by promoting the other's positive self-image, by avoiding offence, encouraging comity, and so on. The negotiation of meaning is also a negotiation of social relations.' [14]

that a text is solely a linguistic phenomenon, comprising a bag of words, the statistical properties of which are adequate to describe it. But the TEI tried very hard to prefer comity over conflict, not only in its organization, which brought together an extraordinarily heterogeneous group of experts, but also in its chief outputs: a set of encoding Guidelines which while supporting specialization did not require any particular specialisation to prevail.

Old orthodoxies do not die easily, and it is interesting to hear how some of the same arguments are still being played out in the somewhat different context of today's DH theorizers. But in our present paper, we simply want to explore the extent to which the TEI's model of text can be adapted to conform to the model of text characterising such fields as stylometry, stylistics, textual analytics, or (to use the current term) ‘distant reading’. We hope also to explore the claim that by so doing we may facilitate the enrichment of that model, and thus facilitate more sophisticated research into textual phenomena across different corpora. And we hope to demonstrate that this is best done by cultivating mutual respect for the widely differing scientific, cultural, and linguistic traditions characterising this cross-European and cross-disciplinary project, that is, by acknowledging a comity of methods as well as languages.

3 The COST Action “Distant Reading for European Literary History”

The context for this work is the EU-funded COST Action “Distant Reading for European Literary History” (CA 16204) a principal deliverable of which will be the European Literary Text Collection (ELTeC). This is a set of comparable corpora for each of at least a dozen European languages, each corpus being a balanced selection of 100 novels from the 19th century, together with metadata situating them in their contexts of production and of reception. It is hoped that the ELTeC will become a reliable basis for comparative work in cross-linguistic data-driven textual analytics, eventually providing an accessible benchmark for a particular written genre of considerable cultural importance across Europe during the period between 1840 and 1920.

Two significant decisions made early on in the planning of the COST Action underlie the work reported here. Firstly, it was agreed that the ELTeC should be delivered in a TEI-encoded format, using a schema developed specifically for the project. Secondly, the design of that encoding scheme, in particular the textual features it makes explicit by means of markup, should be defined as far as possible by the needs of the distant reading research community, rather than any pre-existing notion of textual ontology, to the extent that the needs of that community could be determined. The target audience envisaged includes experts in computational stylistics, in corpus linguistics, and in traditional literary studies as well as more general digital humanists, but is probably best characterized as having major enthusiasm and expertise in the application of statistical methods to literary and linguistic analysis, and only minor interest in the kinds of textual features most TEI projects have tended to focus on.

The work of the Action [2] is carried out in four Working Groups (WGs), whose activities are subject to endorsement and acceptance by a Management Committee, composed of two national representatives from each of the 29 countries currently participating in the Action. The Working Group heads are also members of a smaller 'core' group responsible for day to day management of the Action. WG1 Scholarly Resources is responsible for the work described in this paper; WG2 Methods and Tools is concerned with text analytic techniques and tools; WG3 Literary Theory and History is concerned with applications and implications of those methods and for literary theory ; WG4 Dissemination is responsible for outreach and communication.

The design and construction of the ELTeC is the responsibility of WG1, as noted above. Initially, this work was split into three distinct tasks: First, defining selection criteria (corpus design); second, developing basic encoding methods (both for data and for metadata); and third, defining a suitable workflow for preparation of the corpus. Working papers on each of these topics plus a fourth on theoretical issues of sampling and balance were prepared for discussion and approval by the 30 members of WG1, and remain available from the Working Group's website. Their proposals were ratified by the Management Committee after discussion at two meetings during 2018.

4 The ELTeC Encoding Scheme/s

The encoding requirements for ELTeC were perceived by WG1 to be somewhat different from those of many other TEI projects. Distant Reading methods cover a wide range of computational approaches to literary text analysis, such as authorship attribution, topic modelling, character network analysis, or stylistic analysis but they are rarely concerned with editorial matters such as textual variation, the establishment of an authoritative text, or production of print or online versions of a text. Consequently, the ELTeC encoding scheme was deliberately not intended to represent source documents in all their original complexity of structure or appearance, but rather to make it as simple as possible to access the words of which texts are composed in an informed and predictable way. The goal was neither to duplicate the work of scholarly editors nor to produce (yet another) digital edition of a specific source document. Rather, the encoding scheme was designed in such a way as to ensure that ELTeC texts could be processed by simple minded (but XML-aware) systems primarily concerned with lexis and to make life easier for the developers of such systems.

An important principle following from this latter goal is that ELTeC markup should offer the encoder very little choice, and the software developer very few surprises: the number of tags available is greatly reduced, and their application is tightly constrained. It facilitates processing greatly if access to each part of the

[2] Further information about the Action is available from its website at https://www.distant-reading.net

XML tree can be provided in a uniform and consistent way across multiple ELTeC corpora.

By default, the TEI provides a very rich vocabulary, and many subtly different ways of doing more or less the same thing. TEI encoders have frequently taken full advantage of that to produce texts which vary enormously, both in the subset of XML tags used and in the range of attribute values associated with them. It is tempting, but entirely mistaken, to assume that the allegedly TEI-conformant deliverables from project A will necessarily be marked up in the same way as the allegedly TEI-conformant deliverables from project B [3]. On the contrary, all that 'TEI conformance' really guarantees is that the intended semantics of the markup used by the two projects should be recoverable by reference to a published standard, and are not entirely ad hoc or sui generis. (This may not seem much of an advance, though it is: see further [[5]]).

Following this No Surprises principle, the simplest ELTeC schema (the 'level zero ' schema) provides the bare minimum of tags needed to mark up the typical structure and content of a nineteenth century novel. All preliminary matter other than the title-page and any authorial preface or introduction is discarded; the remainder is marked as a <`div`> of *@type*`titlepage` or `liminal`, within a <`front`> element. Within the <`body`> of a text, the <`div`> element is also used to make explicit its structural organization, with *@type* attribute values `part`, `chapter`, or `letter` only[4]. For our purposes, a 'chapter' is considered to be the smallest subsection of a novel within which paragraphs of text appear directly. Further subdivisions within a chapter (often indicated conventionally by ellipses, dashes, stars etc.) are marked using the <`milestone`> element; larger groupings of <`div`> elements are indicated by <`div`> elements, always of type `part`, whatever their hierarchic level. Headings, at whatever level, are always marked using the <`head`> element when appearing at the start of a <`div`>, and the <`trailer`> element when appearing at the end. Within the <`div`> element, only a very limited number of elements is permitted: specifically, in addition to those already mentioned, <`p`> or <`l`> (verse line). Within these elements we find either plain text, <`hi`> (highlighted), <`pb`> (page break) or <`milestone`> elements. After some debate, the Action's Management Committee agreed that it would be practical to require only this tiny subset of the TEI for all ELTeC texts.

It should be noted that the texts included in an ELTeC corpus may come from different kinds of source. For some language collections, no digital texts of any kind exist: the encoder must start from page images, manually transcribe or put them through OCR, and introduce ELTeC markup from scratch. For others, existing digital texts may already be available: the encoder must research the format used and find a way of converting it to ELTeC. In some cases, a TEI

[3] A large-scale project called MONK (Metadata Offer New Knowledge) demonstrated some of the technical consequences of this for integrated searching of TEI resources: see further http://monk.library.illinois.edu

[4] An exception is made for epistolary novels which contain only the representation of a sequence of letters, with no other significant content: these may be marked as <div type="letter">

version may already exist; in others a project Gutenberg HTML version; in yet others the text may be stored in a database of some kind. Whichever is the case, if it is possible to retain distinctions which the ELTeC scheme permits, this is clearly desirable; perhaps less obviously, it is also necessary to remove distinctions made by the original format which the ELTeC scheme does not permit. This diversity of source material was one motivation for permitting multiple encoding levels in the ELTeC scheme: at level zero, only a bare minimum of markup is required or permitted, while at level 1 a slightly richer (but still minimalist) encoding is also defined. At level 2, further tags again are introduced to support linguistic processing of various kinds, as discussed further below. Down-conversion from a higher to a lower level is always automatically possible, but up-conversion from a lower to a higher level generally requires human intervention or additional processing.

At level 1, the following additional distinctions may be made in an encoding:

- the <label> element may be used for heading-like titles appearing in the middle of a division;
- the <quote> element may be used to distinguish passages such as quotations, epigraphs, stretches of verse, letters etc. which seem to 'float' within the running text;
- the <corr> element may be used to indicate a passage (typically a word or phrase) which is clearly erroneous in the original and which has been editorially corrected;
- the elements <foreign>, <emph>, or <title> are available and should be used in preference to <hi> for passages rendered in a different font or otherwise made visually salient in the source, where an encoder can do so with confidence;
- the element <gap> may be used to indicate where some component of a source (typically an illustration) has been left out of the encoding;
- the elements <note> and <ref> may be used to capture the location and content of authorially supplied footnotes or end-notes; wherever they occur in the source, notes must be collected together in a <div type="notes"> within a <back> element.

For those already familiar with the TEI, this list of elements may seem distressingly small. It lacks entirely some elements which every TEI introductory course regards as indispensable (no <list> or <item>; no <choice> or <abbr>; no <name> or <date>...) and tolerates some practices bordering on tag abuse. For example, all the components of a title page are marked as <p> since no specialised elements (<titlePage>, <docImprint> etc.) are available. In the absence of specialised but culture-specific features (for example, publisher name, imprint, imprimatur, etc.) the encoding identifies only fundamental textual features common to every kind of text. Nevertheless, we believe that the set of concepts it supports overlaps well with the set of textual features which almost any existing digital transcription will seek to preserve in some form or another. This may explain both why the majority of the texts so far collected in the ELTeC have been encoded at level 1 rather than level0, and also the speed with which the collection is growing.

ELTeC level 1 is intended to facilitate a richer and better-informed distant reading of a text than a transcription of its verbal content alone would permit. ELTeC level 2 is partly intended to provide a consistent and TEI-conformant way of representing the results of such readings, in partiocular those concerned with linguistic annotation. Its primary goal is to represent in a standard way additional layers of annotation of particular importance to distant reading applications such as stylometry or topic modelling. Enrichment of each lexical token to indicate its morpho-syntactic category (POS) or its lemma, and identification of tokens which refer to named entities are both well within the scope of existing text processing techniques, and are also routinely used in distant reading applications. The challenge is that the input and the output formats typically used by such tools are rarely XML-based, and seem superficially to have a model of text quite different from that of the ‘ordered hierarchy of content objects’ in terms of which the TEI community traditionally operates. For many in the distant reading community (it seems) a text is little more than a sequence of tokens, mostly corresponding with orthographically-defined words, though there is some variability in the principles underlying the process of tokenisation, for example in the modelling of clitics, compound forms, etc. Each token has a number of properties, which might include such attributes as its part of speech, its lemma, or its position in the sequence of tokens making up the document. Information about a token which in an XML model would be properties of some higher level construct such as its status as dialogue, quoted matter, emphasis, etc. is occasionally considered as well, but is typically modelled as an additional property of the token.

If a community is defined by its tools, it would appear therefore that the distant reading community has not fully embraced the notion of XML as anything other than a rather verbose archival format. However, communities are not defined solely by their tools : by seeking a way of reconciling these differing views of what text really is in a spirit of comity we hope to demonstrate that there are advantages both for the distant reader or stylometrician and for the literary analyst or textual editor.

At ELTeC level2, all existing elements are retained and two new elements <s> and <w> are introduced to support segmentation of running text into sentence-like and word-like sequences respectively. Individual tokens are marked using the <w> element, and decorated with one or more of the TEI-defined linguistic attributes *@pos*, *@lemma*, and *@join*. Both words and punctuation marks are considered to be ‘tokens’ in this sense, although the TEI suggests distinguishing the two cases using <w> and <pc> respectively. The <s> (segment) element is used to provide an end-to-end tessellating segmentation of the whole sequence of <w> elements, based on orthographic form. This provides a convenient extension of the existing text-body-div hierarchy within which tokens are located.

The elements <p>, <head>, and <l> (which contain just text at levels 0 and 1) at level 2 can contain a sequence of <s> elements. Empty elements <gap>, <milestone>, <pb> or <ref> are also permitted within text content at any point, but these are disregarded when segmentation is carried out. Each <s> element can contain a sequence of <w> elements, either directly, or wrapped in one of the sub-paragraph elements <corr>, <emph>, <foreign>, <hi>, <label>, <title>. To this list we might add the element <rs> (referring string), provided by the TEI for the encoding of any form of entity name, such as a Named Entity Recognition procedure might produce.

This approach implies that <w> elements may appear at two levels in the hierarchy which may upset some software; it also implies that <w> elements must be properly contained within one of these elements, without overlap. If either issue proves to be a major stumbling block, an alternative would be to remove the tags demarcating these sub-paragraph elements, indicating their semantics instead by additional attribute values on the <w> elements they contain.

This TEI XML format is equally applicable to the production of training data for applications using machine learning techniques and to the outputs of such systems. However, since such machine learning applications typically operate on text content in a tabular format only, XSLT filters which transform (or generate) the XML markup discussed here from such tabular formats without loss of information are envisaged. At the time of writing, however, Working Group 2 has yet to put this proposed architecture to the test.

5 ELTeC metadata and corpus design

Like every other TEI document, every ELTeC text has a TEI Header, though its organization and content alike are constrained much more tightly than is common TEI praxis, for the reasons already mentioned. The structure of an ELTeC Header is the same no matter what level of encoding applies to the text. It provides minimal bibliographic information about the encoded text and its source, sufficient to identify the text and its author, in a fixed and consistent format. It is assumed that if more detailed bibliographic information is required, for example about the author or work encoded, this is better obtained from standard authority files; to that end a VIAF code may be associated with them.

As noted above, ELTeC texts may be derived from many sources, each of which should be documented correctly in the header's <sourceDesc> element. After some debate, a common set of practices has been identified to distinguish (for example) ELTeC texts derived directly from a print source from those derived from a digital source, itself derived from a known print source, and to provide information about each source. In the following example, the source of the ELTeC version is a pre-existing digital edition provided by Project Gutenberg , but the source description also provides information about the first print edition of the work concerned.

```
<bibl type="digitalSource">
 <title>Project Gutenberg EBook A engomadeira de Almada Negreiros</title>
 <ref target="http://www.gutenberg.org/ebooks/23879"/>
</bibl>
<bibl type="firstEdition">
 <title>A engomadeira</title>
 <author>José de Almada Negreiros</author>
 <publisher>Typographia Monteiro &amp; Cardoso</publisher>
 <date>1917</date>
</bibl>
```

In most cases, the ELTeC text will correspond with the first edition of a work in book form; but even where this is not the case, or where information about the precise source used is not available, minimal information about that first edition should also be provided in order to place the work in its original temporal context.

As with other TEI conformant documents, beside the mandatory file description, the TEI Header of every ELTeC text contains a publication statement which specifies its licensing conditions (all ELTeC documents are licensed CC-BY); an encoding statement specifying the level of encoding used; and a revision description containing versioning information. The TEI Header is also used to provide metadata describing the associated text in a standardized form; this is held in the <`profileDesc`> element which must specify the languages used by the text, may optionally include a <`textClass`> element containing any culture-specific keywords considered useful to describe the text, and must contain a <`textDesc`> element which documents the text's status with respect to selection criteria discussed below.

One of the knottier problems or (to be positive) more distinctive features of an ELTeC language collection is that it is not intended to be an *ad hoc* accidentally constructed corpus but a designed one. Its composition is determined not by the happenstance of whatever we can get our hands on, but is instead defensible, at least in theory, as a principled and representative selection.

The big question is, of course, representative of *what*.

It would be nice to say that it represents the production of novels in a specific language in 19th century Europe. WG1 has working definitions for both "novels" and "Europe" which we do not discuss further here, though both are clearly problematic terms. It is hoped that the ELTeC will provide data for an empirical discussion of such terms, feeding into the work of WG3 on literary theory and terminology.

But we cannot make that claim without any data about the population we are claiming to represent -- which is hard to come by for many of the languages concerned. We know about the novels which we know about, which tend to be the ones that national libraries or equivalent cultural heritage institutions have chosen to preserve, which publishers over time have been able to sell, and which lecturers in literary studies have chosen to teach. More ephemeral titles may have been collected (for example by a copyright library); but equally well may have been discarded or even suppressed as unworthy of inclusion in the national patrimony. Titles and authors alike can go in and out of fashion. But how can we express opinions about changes in the nature of the published novel if the sample on which we base those opinions is wildly different in composition from the actual population? If our data leads us to assert that novels in a given language are never written by women, or are never of fewer than 100,000 words is this simply because no female authors happen to have been preserved, or because short novels were routinely discarded from the collection? Or, on the other hand, does this actually indicate something fundamental, a characteristic of the population we are investigating? This matters particularly for ELTeC, one of the goals of which is precisely to facilitate cross-language comparisons.

This problem of representativeness is of course one which every corpus linguist has to face, and discussions of its implications are easy to find in the literature [5]

Our approach is to sidestep the impossibility of representing an unknown (and sometimes unknowable) population by attempting instead to represent the range of possible variation in the values of a predefined set of variables, each corresponding with a more or less objective category of information available for all members of the population. To take a trivial example, every novel can be characterised as short, medium, or long; there is no possible fourth value for this category unless we revise our definition of length (elastic? unknown? instantaneous?). So, as a working hypothesis, we might say that a corpus in which roughly a third of the titles are short, a third are long, and a third are medium will represent the variation possible for this category. If we apply this principle uniformly across all our corpora, we can reliably investigate (for example) cross language variation in some other observable phenomenon (say a fondness for syntactically complex sentences) with respect to length. But note that we have made absolutely no claim about whether novel length in the underlying population is also divided in this way.

The decade in which a novel first appears in book form is a similarly objectively characteristic, which in principle we can determine for every member of the population. We can also classify every title according to the actual sex of their author (with values such as female, male, mixed, unknown). And we can likewise classify a title in terms of its staying power or persistence by looking at the number of times it has been reprinted since its first appearance. We suggest that texts which have been frequently reprinted over a long period may reasonably be considered ‘canonical’ in some sense of that vexed term. The goal of our corpus balancing exercise is to ensure more or less equal time for each possible value for each of these four categories -- size, decade, authorSex, and canonicity.

[5] Some notable examples include [3]; [11]; [4]

Ideally, each corpus should have equivalent numbers not just for each value, but for each combination of values: so, for example, looking at the third of all titles which are characterised as "short", there should be roughly equal numbers for each decade of first appearance, roughly equal numbers by male and female authors, and so on. This may however be a council of perfection. It is already apparent that for some languages, it is very difficult to find any texts at all within some time periods, or by female authors. Similarly, our definition of "short" (10-50 thousand words), "medium" (50-100 thousand words) and "long" (over 100 thousand words) though objective and easy to validate, assumes that there will be enough novels of a given length in the underlying population for us to extract a balanced sample; but in some languages it may be that the distribution of lengths across the population is entirely different. We cannot tell whether (for example) the absence of any "long" novels at all in Czech, Serbian, or Norwegian is characteristic of those languages, or an artefact of the selection process so far. Another difficulty is that our corpus design deliberately seeks to include some forgotten or marginal works along with well-known canonical texts: this is relatively easy for traditions such as English, French, or German where copyright laws have led to the maintenance and documentation of large national collections, but less so for other less well documented languages. (For some initial data, see the summary page at https://distantreading.github.io/ELTeC/)

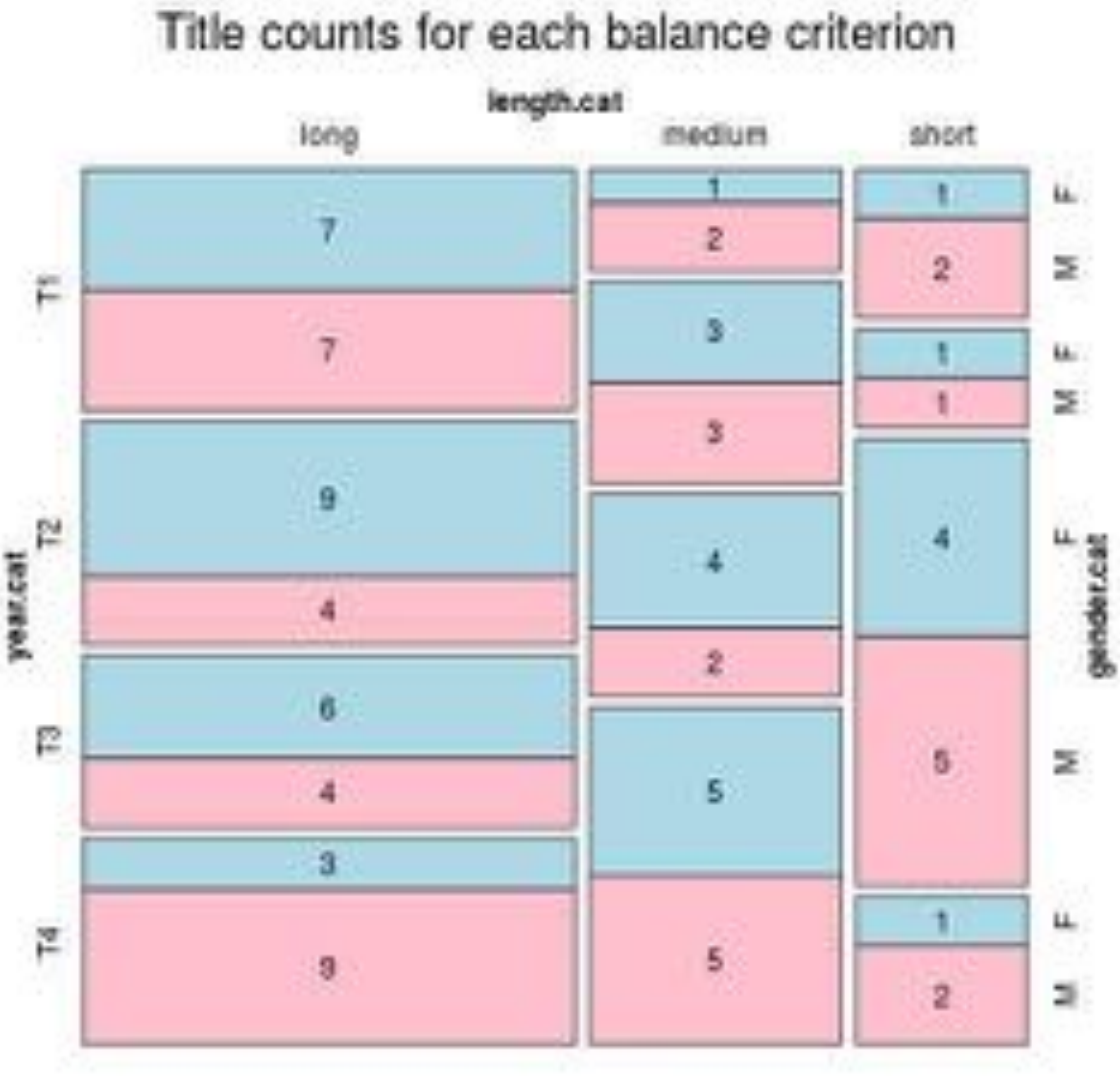
To encode these balance criteria in the TEI Header in as direct and accessible a manner as possible, we have chosen to re-purpose the little-used <`textDesc`> element, originally provided by the TEI as a wrapper for a set of so-called situational parameters proposed by corpus linguists as a way of objectively characterizing linguistic production [6] In our case, we replace the TEI's suggested vocabulary for these parameters with a vocabulary representing our four criteria, expressed as new non-TEI elements in the ELTeC namespace. These elements (<`eltec:sex`>, <`eltec:size`>, <`eltec:canonicity`>, and <`eltec:timeSlot`>) are required by the ELTeC schemas and have an attribute *@key* which supplies a coded value for the criterion concerned taken from a predefined closed list. So, for example, a long (over 100,000 words) novel by a female author first published between 1881 and 1900 but only infrequently reprinted thereafter might have a text description like the following:

```
<textDesc
   xmlns:eltec="http://distant-reading.net/ns">
 <eltec:authorGender key="F"/>
 <eltec:canonicity key="low"/>
 <eltec:size key="long"/>
 <eltec:timeSlot key="T3"/>
</textDesc>
```

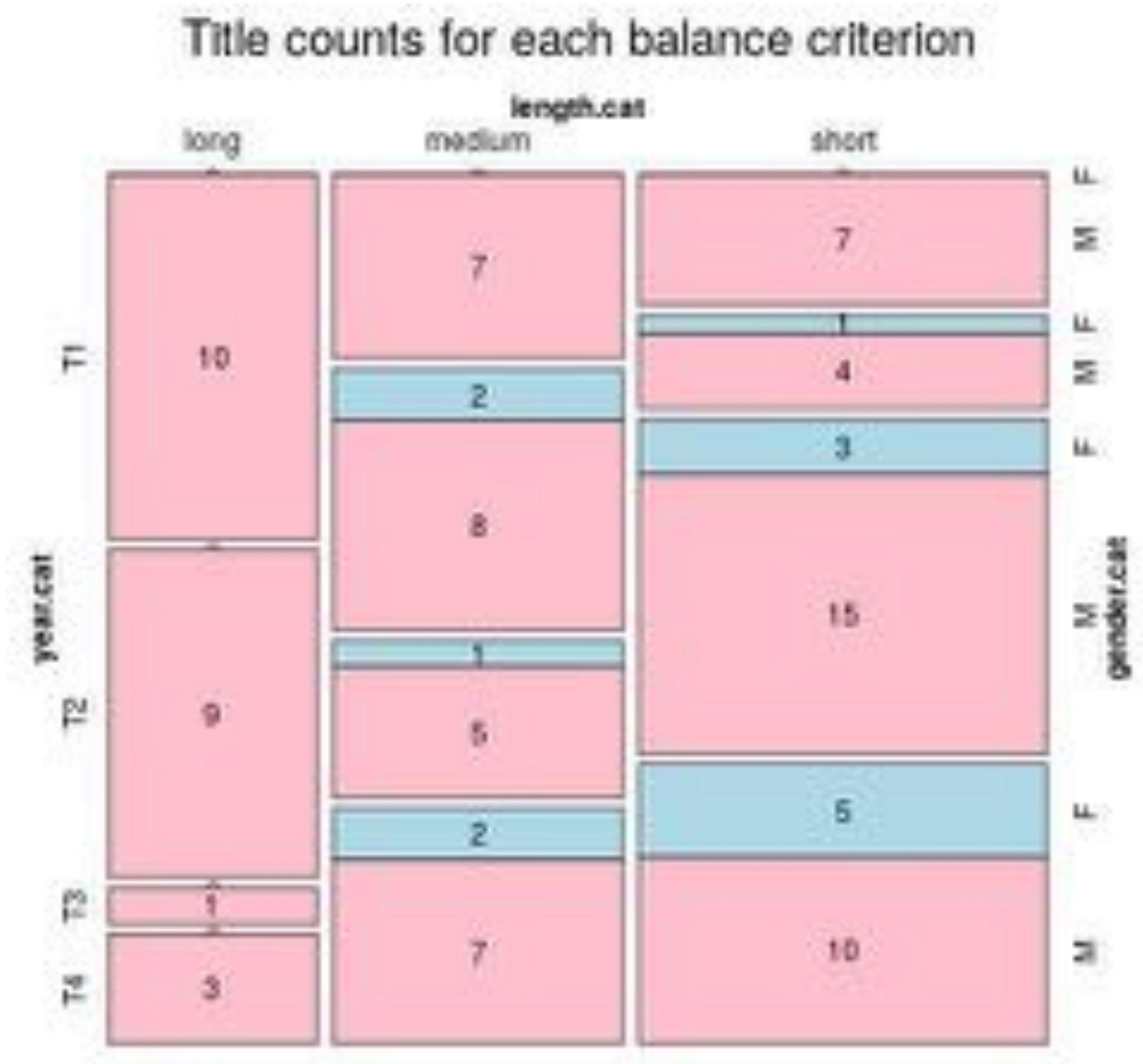
When complete, this information can be used to select subcorpora from the collection as a whole, thus permitting more delicate cross-linguistic comparisons: for example between the lexis of male and female writers, or between the

[6] The <`textDesc`> element is discussed in section 15.2.1 of the TEI *Guidelines* (https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAHTD).

stylistic features typically associated with long or short texts. During the construction phase, these coded values also make it easy to monitor the emerging composition of the corpus, for example to detect whether or not the ratio of male to female writers is consistent across different time periods, by means of a simple visualisation like the following



This ‘mosaic plot’ for the current state of the English corpus (90 texts) shows that there are roughly as many female (blue) as male (pink) writers across the board, but that there is a preponderance of long texts and of titles published in time slot 3.



For comparison, the same plot for the current state of the Hungarian corpus (100 texts) shows significantly fewer female writers, and a higher proportion of short texts. Whether these variations are an artefact of the sampling process or represent differences in the underlying population is precisely one of the research questions which our approach requires us to address.

6 Chaining ODDs

The TEI's ODD (One Document Does it all) system [[12]] is widely used as a means of customizing the TEI and documenting the customization in a standard way. When only a single ODD customization is used across a project, there is a natural tendency to produce broadly permissive schemas, to allow for the inevitable variation of requirements when material of different kinds are to be processed in an integrated collection. But this prevents the encoder from taking full advantage of the ability of an XML schema to check that particular documents conform to predefined rules, unless they are willing greatly to increase the complexity of their work flow. A better approach, pioneered by the Deutsch Textarchiv [10], has been the use of a technique known as ODD chaining [5] Here, a project first defines a base ODD which selects all the TEI components considered to be useful anywhere and then uses this as the basis for smaller, more constraining, ODDs which select from the base only the components (or other rules) specific to a subset of the project's documentary universe. For example, an archive may have identified a common set of metadata it wishes to document across all of its holdings but also have particular metadata requirements for print and manuscript sources respectively. Simply defining two different ODDs, one for print and one for manuscript when many other components apply to either kind of source opens the door to redundant duplication and the risk of inconsistency. The ODD chaining approach requires definition of a base ODD which contains the union of the components needed for these two different ODDs, constructed as an appropriate selection from the full range of TEI components. The ODDs for print and manuscript are then defined as further specialisations or customizations of the base, ensuring thereby that the common components are used in a consistent manner, but preserving comity by allowing equal status to the two specialised schemas.

In the ELTeC project, we begin by defining an ODD which selects from the TEI all the components used by any ELTeC schema at any level. This ODD also contains documentation and specifies usage constraints applicable across every schema. This base ODD is then processed using the TEI standard `odd2odd` stylesheet to produce a standalone set of TEI specifications which we call `eltec-library`. Three different ODDs, eltec-0, eltec-1, and eltec-2 then derive specific schemas and documentation for each of the three ELTeC levels, using this library of specifications as a base rather than the whole of the TEI. This enables us to customize the TEI across the whole project, while at the same time respecting three different views of the resulting encoding standard. As with other ODDs, we are then able to produce documentation and formal schemas which reflect exactly the scope of each encoding level.

The ODD sources and their outputs are maintained on GitHub and are also being published on Zenodo along with the ELTeC texts. [7]

7 State of play and future work

The ELTeC is still very much a work in progress, and hence we cannot report that our design goals have been achieved with any plausibility. An initial release of the collection is due on Zenodo in September 2019, and we expect several future releases before the target of 100 texts per language is reached. The corpora are also maintained as a collection of publicly visible GitHub repositories, as noted above.

[7] The GitHub repository for the ELTeC collection is found at https://github.com/COST-ELTeC/ : the Zenodo community within which it is being published lives at: https://zenodo.org/communities/eltec,

As well as continuing to expand the collection, and continuing to fine-tune its composition, we hope to improve the consistency and reliability of the metadata associated with each text, as far as possible automatically. For example, we have developed two complementary methods of automatically counting the number of reprints for each title, one by screen scraping from WorldCat, and the other by processing data from a Z39.50 server where this is available. These methods should provide more reliable data than has hitherto been available for the 'canonicity' criterion mentioned above.

The main area of future work we anticipate is however in the testing of the proposed ELTeC level 2 encoding and an evaluation of its usefulness. At a technical level, this may necessitate some changes in the existing markup scheme, but of perhaps more interest is the extent to which its availability will exemplify the virtue of striving for comity amongst the many ways in which TEI XML markup can be applied.

[1] References

[2] Aston, Guy (1988) Learning Comity: An Approach to the Description and Pedagogy of Interactional Speech (Testi e discorsi: Strumenti linguistici e letterari, vol 9) Bologna: CLUEB

[3] Biber, Douglas (1993). "Representativeness in Corpus Design". In: Literary and Linguistic Computing (8), pp. 243–257.

[4] Bode, Katherine (2018). A World of Fiction - Digital Collections and the Future of Literary History. eng. University of Michigan Press.

[5] Burnard, Lou (2016) ODD Chaining for Beginners. Available from http://teic.github.io/PDF/howtoChain.pdf

[6] Burnard, Lou (2019) "What is TEI Conformance, and why should you care?". In: Journal of the Text Encoding Initiative, Issue 12. https://journals.openedition.org/jtei/1777

[7] Caton, Paul (2013). "On the term text in digital humanities". In: Literary and Linguistic Computing 28.2, pp. 209–220.

[8] De Rose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear (2002). "What is Text, Really?" In: Journal of Computing in Higher Education I(2), pp. 3–26.

[9] Gavin, Michael (2019) "How to think about EEBO". In: Textual Cultures Vol 11, no 1-2 (2017). https://doi.org/10.14434/textual.v11i1-2.23570

[10] Haaf, Susanne and Christian Thomas (2016) "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the

Format” In: Journal of the Text Encoding Initiative, Issue 10. https://journals.openedition.org/jtei/1650

[11] Lüdeling, Anke (2011). “Corpora in Linguistics. Sampling and Annotation”. In: Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Ed. by Karl Grandin. Vol. 147. Nobel Symposium 147. New York:

[12] Rahtz, Sebastian, and Lou Burnard (2013) “Reviewing the TEI ODD System”. In Proceedings of the 2013 ACM Symposium on Document Engineering. DocEng ’13. ACM, 2013. http://doi.acm.org/10.1145/2494266.2494321

[13] van Zundert, Joris and Tara L. Andrews (2017). “Qu’est-ce qu’un texte numérique? A new rationale for the digital representation of text”. In: Digital Scholarship in the Humanities 32, pp. 78–88.

[14] Widdowson, Henry (1990) Aspects of Language Teaching. OUP.

Creating high-quality print from TEI documents

Even in the age of digital editions a need for high-quality print versions of TEI documents remains, be it a print version of a single document or a book edition of a complete or partial digital edition.

Of course, there is XSL-FO but the typesetting quality obtained by the freely available Apache FOP does not live up to very high demands. Obviously, one can yield to one of the commercial XSL-FO engines which do a rather impressive job. However, even the best XSL-FO engines are bound by the inherent limitations of XSL-FO which are crucial for critical editions, e.g. for formatting a complex text-critical apparatus.

Fortunately, there has been a freely available typesetting system since the early 1980s whose quality is widely acknowledged, namely LaTeX and the underlying TeX. The use of LaTeX for creating scholarly editions is therefore well-established (cf. the reledmac package and its predecessors). Likewise, LaTeX has been utilised before for typesetting TEI documents, e.g. within the XSL stylesheets for the TEI framework. However, using XSLT for transforming TEI documents to various formats has drawbacks, especially a rather high degree of redundancy and the consequential pain of maintenance.

Luckily enough, the TEI Processing Model and ODD come to our rescue. They provide means for adding LaTeX to the output formats without too much hassle. Using the TEI Publisher Libraries complex requirements can be easily realised within the TEI ODD in a descriptive, standardised way.

This paper demonstrates how the portal of the Law Sources Foundation of the Swiss Lawyers Society utilises the TEI Processing Model and LaTeX for preparing high-quality print versions. We will also propose two simple extensions to the TEI Processing Model, `pb-template` and `pb-behaviour`, which make it easier to generate arbitrary LaTeX from within the ODD using a straightforward template syntax As our implementation proves, with those minimal extensions the Processing Model becomes powerful and generic enough to cover complex, high-quality print.

Refining the Current Teaching Methodology of the TEI through the Analysis of Server Logs

Luis Meneses, Electronic Textual Cultures Lab, University of Victoria, ldmm@uvic.ca
Jonathan Martin, King's College London, jonathan.d.martin@kcl.ac.uk

Keywords: server logs, teaching methodology, TEI Guidelines

We believe that the next step in the evolution of the TEI is developing training materials, which aligns with the emphasis that has been placed lately on the pedagogy and practice of the Guidelines –which was the main theme of the 2017 conference ("TEI Victoria 2017" 2019). The materials for learning the TEI Guidelines are still in early stages –consisting primarily of past project documentation, the TEI's own introductory materials, online tutorials, and collected examples –which leads to skills being acquired in unsystematic ways (Jakacki 2016). Additionally, the Guidelines have evolved and become more rigorous and theoretical –making some of these training materials overwhelming and unpractical for newcomers who might not be familiar with text encoding. As proponents of the TEI Guidelines, we have an obligation to develop equally adequate and appropriate training materials for new learners.

The current TEI Infrastructure consists of a set of servers and services (Cayless et al. 2018), allowing the Guidelines to be primarily accessed online. A server log is a file automatically created and maintained by a server consisting of a list of activities it performed. For our purposes, a statistical analysis of server logs can be used to examine Web traffic patterns. In this abstract, we propose to analyze the TEI server logs in order to offer suggestions to refine the teaching methodology based on what parts of the Guidelines are more frequently accessed –while also considering what is not accessed often. Additionally, customizations of the Guidelines exist that aim to meet the majority of the needs of TEI user community –for example, TEI Lite ("TEI Lite – TEI: Text Encoding Initiative" 2019). However, formal justifications for which elements are included and excluded in them do not exist. This proposal interrogates and theorizes how we might present the TEI Guidelines as better teaching materials and aims to foster the development of skills and activities of future scholars.

References

Cayless, Hugh, James Cummings, Martin Holmes, Peter Stadler, and Magdalena Turska. 2018. "TEI Technical Infrastructure." presented at the TEI 2018: 18th annual Conference and Members Meeting of the Text Encoding Initiative Consortium, Tokyo, Japan, September 9.

Jakacki, Diane. 2016. "How We Teach? Digital Humanities Pedagogy in an Imperfect World." *Diane Jakacki*. June 5. http://dianejakacki.net/how-we-teach-digital-humanities-pedagogy-in-an-imperfect-world/.

"TEI Lite – TEI: Text Encoding Initiative." 2019. Accessed April 22. https://tei-c.org/guidelines/customization/lite/.

"TEI Victoria 2017." 2019. Accessed April 22. https://hcmc.uvic.ca/tei2017/.

Author's Biographies

Dr. Luis Meneses is a Postdoctoral Fellow at the University of Victoria. He is a Fulbright scholar, and currently serves on the Board of the TEI Consortium and on the IEEE Technical Committee on Digital Libraries. His research interests include digital humanities, digital libraries, information retrieval and human-computer interaction.

Jonathan Martin is a visiting graduate researcher and a member of the Ph.D. program in Digital Humanities at King's College London. The focus of his doctoral work is an ethnography of digital humanists, which he is currently undertaking in the Electronic Textual Cultures Laboratory at the University of Victoria. The purpose of this work is to explore digital humanities scholarship as a practice, as a way of making and collaborating, and as a part of a larger academic culture.

How we tripled our encoding speed in the Digital Victorian Periodical Project

Kaitlyn Fralick, Kailey Fukushima, Martin Holmes, and Sarah Karlson

ABSTRACT

The Digital Victorian Periodical Poetry (DVPP) project is a SSHRC-funded digital humanities project based at the University of Victoria. With the guidance of principal investigator Dr. Alison Chapman, the DVPP team is creating a digital index of British periodical poetry from the long nineteenth century. In addition to uncovering periodical poems, writing descriptive metadata, and compiling prosopographical research, we are currently using TEI and CSS to encode a statistically-representative sample of indexed poems, looking for quantitative evidence of literary change over time. Such an endeavour requires a large, robust dataset covering a range of periodicals throughout the period.

At the time of writing, there are more than 13,000 poems in the database, and we expect that total to reach 20,000. Of these, around 2,000 will be encoded, focusing on the decade years (1820, 1830, 1840, and so on).

In this presentation, we will showcase the various strategies and tools we have used to speed up our encoding process. We combine simple tricks like keyboard shortcuts with more sophisticated processes to minimize drudgery and increase accuracy. Among the more interesting techniques are:

- Auto-tagging of a complete poem in lines and linegroups using a Schematron QuickFix;
- Use of advanced CSS selectors in the rendition/@selector attribute to reduce encoding clutter in the poem itself;
- A keyboard shortcut to tag rhymes which detects whether the tagged text is a masculine or feminine rhyme and provides the appropriate attribute value;
- Auto-detection of cases where a new line-end rhymes with a previously-encoded rhyme, and should, therefore, be labelled to match it, leveraging our growing dataset of nearly 30,000 rhymes;
- Instant access to to a rendering of the poem which provides a visualization of the rhyme structure, auto-detection of anaphora, epistrophe and other refrain-like forms, and other diagnostic feedback.

INDEX

Keywords: encoding tools, encoding strategies, TEI simplification, TEI environments and infrastructures, visualization

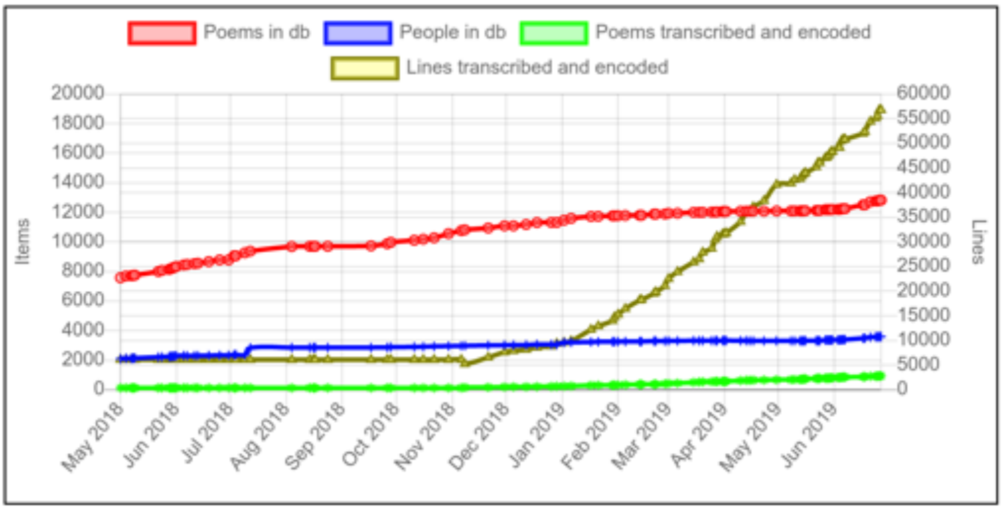
1. Introduction

1 The Digital Victorian Periodical Poetry (DVPP) project is a SSHRC-funded digital humanities project based at the University of Victoria. With the guidance of principal investigator Dr. Alison Chapman, the DVPP team is creating a digital index of British periodical poetry from the long nineteenth century. In addition to uncovering periodical poems, writing descriptive metadata, and compiling prosopographical research, we are currently using TEI and CSS to encode a statistically-representative sample of indexed poems, looking for quantitative evidence of literary change over time. Such an endeavour requires a large and robust dataset covering a range of periodicals throughout the period.

2 At the time of writing, there are over 13,000 poems in the database, and we expect that total to reach 20,000. Of these, around 2,000 will be encoded, focusing on the decade years (1820, 1830, 1840 and so on).

3 When we initially received funding for this project, we were quite confident that we would be able to accomplish the collection of page-images and metadata across the period, because we had been doing that work for some years and had a clear idea of the time and resources required. However, a relatively small number of poems had been encoded, and the encoding time required had proved to be significant. Our first pass of encoding goes beyond simple transcription and tagging of lines and line-groups to include rhyme-schemes and rhyme types, refrains and similar devices such as anaphora and epistrophe, and detailed description of typographical style and layout using CSS. These are all features that will form part of the analytic work to be done when the encoding is complete.

Figure 1. Encoding speed improvements (yellow line with triangle markers).



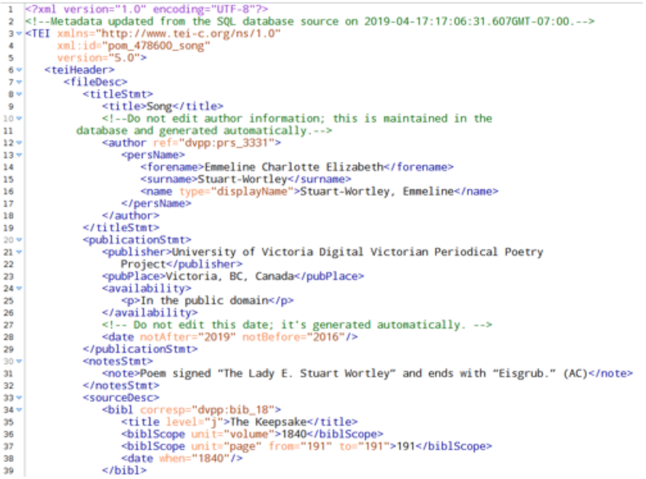
4 Initial projections from our project diagnostics suggested that we would have some difficulty in accomplishing the tagging task within the time available, so beginning towards the end of 2018, the encoding team (Fralick, Fukushima, Holmes, and Karlson) began a concerted effort to develop tools and techniques to make our encoding faster and more effective. The results can be seen in the graph in figure 1, taken from our project diagnostics. In this presentation, we will describe

and demonstrate some of the techniques we deployed to make this improvement possible. We hope these examples, illustrations, and suggestions will be useful to other projects endeavouring to make their encoding process more accurate and efficient.

2. Automatic headers and OCR

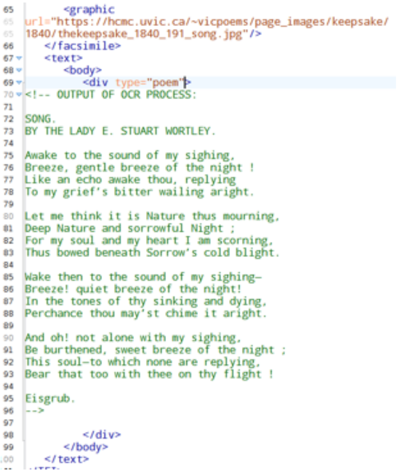
5 First of all, we eliminated all the work of setting up the XML file, completing the core metadata, and doing the transcription. Our large database index of poetry includes metadata on all the poems, along with links to the page-images in which they appear. A build process is able to export the SQL database as XML, process it into TEI files, and initiate an OCR process on each file, storing the results in the form of a comment in the body of the TEI file. This removes the bulk of the transcription work, and relieves the encoder of any responsibility for the metadata. The quality of the OCR is variable, but in almost all cases, proofing and correcting it is much faster than transcribing manually (figure 2 and figure 3).

Figure 2. The TEI Header generated automatically from the database.



```xml
<?xml version="1.0" encoding="UTF-8"?>
<!--Metadata updated from the SQL database source on 2019-04-17:17:06:31.607GMT-07:00.-->
<TEI xmlns="http://www.tei-c.org/ns/1.0"
     xml:id="pom_478600_song"
     version="5.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Song</title>
            <!--Do not edit author information; this is maintained in the
        database and generated automatically.-->
            <author ref="dvpp:prs_3331">
               <persName>
                  <forename>Emmeline Charlotte Elizabeth</forename>
                  <surname>Stuart-Wortley</surname>
                  <name type="displayName">Stuart-Wortley, Emmeline</name>
               </persName>
            </author>
         </titleStmt>
         <publicationStmt>
            <publisher>University of Victoria Digital Victorian Periodical Poetry
               Project</publisher>
            <pubPlace>Victoria, BC, Canada</pubPlace>
            <availability>
               <p>In the public domain</p>
            </availability>
            <!-- Do not edit this date; it's generated automatically. -->
            <date notAfter="2019" notBefore="2016"/>
         </publicationStmt>
         <notesStmt>
            <note>Poem signed “The Lady E. Stuart Wortley” and ends with “Eisgrub.” (AC)</note>
         </notesStmt>
         <sourceDesc>
            <bibl corresp="dvpp:bib_18">
               <title level="j">The Keepsake</title>
               <biblScope unit="volume">1840</biblScope>
               <biblScope unit="page" from="191" to="191">191</biblScope>
               <date when="1840"/>
            </bibl>
```

Figure 3. The results of OCR, stored as a comment in the XML.

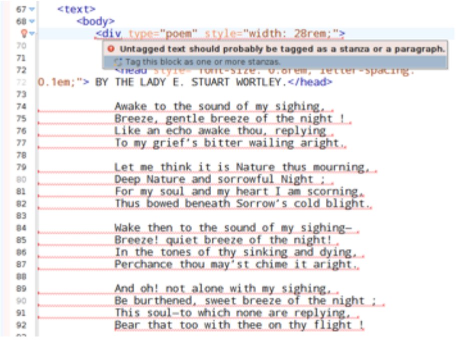


```
        <graphic
url="https://hcmc.uvic.ca/~vicpoems/page_images/keepsake/
1840/thekeepsake_1840_191_song.jpg"/>
      </facsimile>
      <text>
        <body>
          <div type="poem">
<!-- OUTPUT OF OCR PROCESS:

SONG.
BY THE LADY E. STUART WORTLEY.

Awake to the sound of my sighing,
Breeze, gentle breeze of the night !
Like an echo awake thou, replying
To my grief's bitter wailing aright.

Let me think it is Nature thus mourning,
Deep Nature and sorrowful Night ;
For my soul and my heart I am scorning,
Thus bowed beneath Sorrow's cold blight.

Wake then to the sound of my sighing—
Breeze! quiet breeze of the night!
In the tones of thy sinking and dying,
Perchance thou may'st chime it aright.

And oh! not alone with my sighing,
Be burthened, sweet breeze of the night ;
This soul—to which none are replying,
Bear that too with thee on thy flight !

Eisgrub.
-->

          </div>
        </body>
      </text>
```

3. Auto-tagging of an entire poem

6 After correcting the results of the OCR against the original page-image, the next task for the encoder is to tag the structure of the poem. Taking advantage of the support for Schematron QuickFix in the Oxygen XML Editor, we are able to completely automate the tagging of lines and line-groups; the encoder simply has to copy/paste the text version of the poem into a `<div>`, right-click, and choose a quick-fix, and the poem is automatically tagged with `<lg>` and `<l>` elements (figure 4 and figure 5).

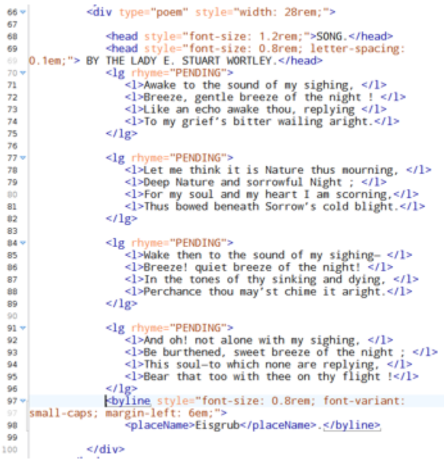
Figure 4. A Schematron QuickFix for auto-tagging an entire poem.



```
    <text>
      <body>
        <div type="poem" style="width: 28rem;">


          <head style="font-size: 0.8rem; letter-spacing:
0.1em;"> BY THE LADY E. STUART WORTLEY.</head>

                Awake to the sound of my sighing,
                Breeze, gentle breeze of the night !
                Like an echo awake thou, replying
                To my grief's bitter wailing aright.

                Let me think it is Nature thus mourning,
                Deep Nature and sorrowful Night ;
                For my soul and my heart I am scorning,
                Thus bowed beneath Sorrow's cold blight.

                Wake then to the sound of my sighing—
                Breeze! quiet breeze of the night!
                In the tones of thy sinking and dying,
                Perchance thou may'st chime it aright.

                And oh! not alone with my sighing,
                Be burthened, sweet breeze of the night ;
                This soul—to which none are replying,
                Bear that too with thee on thy flight !
```

Untagged text should probably be tagged as a stanza or a paragraph.

Figure 5. The result of applying the Schematron QuickFix.

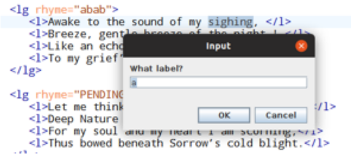


```
        <div type="poem" style="width: 28rem;">

           <head style="font-size: 1.2rem;">SONG.</head>
           <head style="font-size: 0.8rem; letter-spacing:
0.1em;"> BY THE LADY E. STUART WORTLEY.</head>
           <lg rhyme="PENDING">
              <l>Awake to the sound of my sighing, </l>
              <l>Breeze, gentle breeze of the night ! </l>
              <l>Like an echo awake thou, replying </l>
              <l>To my grief’s bitter wailing aright.</l>
           </lg>

           <lg rhyme="PENDING">
              <l>Let me think it is Nature thus mourning, </l>
              <l>Deep Nature and sorrowful Night ; </l>
              <l>For my soul and my heart I am scorning,</l>
              <l>Thus bowed beneath Sorrow’s cold blight.</l>
           </lg>

           <lg rhyme="PENDING">
              <l>Wake then to the sound of my sighing– </l>
              <l>Breeze! quiet breeze of the night! </l>
              <l>In the tones of thy sinking and dying, </l>
              <l>Perchance thou may’st chime it aright.</l>
           </lg>

           <lg rhyme="PENDING">
              <l>And oh! not alone with my sighing, </l>
              <l>Be burthened, sweet breeze of the night ; </l>
              <l>This soul–to which none are replying, </l>
              <l>Bear that too with thee on thy flight !</l>
           </lg>
           <byline style="font-size: 0.8rem; font-variant:
small-caps; margin-left: 6em;">
              <placeName>Eisgrub</placeName>.</byline>

        </div>
```

4. Keystroke shortcuts with special sauce

7 Among the more straightforward tricks are obviously keystroke shortcuts for inserting special characters and XML tags. In addition to the basics (`Alt + QUOTE` to insert a curly apostrophe, `Alt + M` to insert an em dash) we have also been able to take advantage of the fact that Oxygen allows the use of XPath in keystroke shortcuts to provide a useful time-saver when tagging rhyme. `Control + Alt + 0` will wrap a `<rhyme>` tag around selected text, but it does more than that; it uses XPath regular expressions[1] to analyze the contents of the tag to determine whether the rhyme being tagged is most likely masculine (one syllable) or feminine (two or more syllables), and applies the appropriate attribute value to the rhyme element. Although there are of course other varieties of rhyme, in this collection masculine rhymes constitute 84% of the total and feminine rhymes more than 9%. This little trick therefore provides an accurate rhyme tag more than 90% of the time (figure 6 and figure 7).

Figure 6. The encoder selects text, hits the shortcut and provides a label attribute.



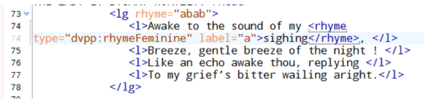
```
<lg rhyme="abab">
   <l>Awake to the sound of my sighing, </l>
   <l>Breeze, gent
   <l>Like an echo
   <l>To my grief’
</lg>

<lg rhyme="PENDING
   <l>Let me think                                 </l>
   <l>Deep Nature
   <l>For my soul
   <l>Thus bowed beneath Sorrow’s cold blight.</l>
```

Input

What label?

a

Figure 7. The shortcut automatically detects the rhyme type as feminine.



```
            <lg rhyme="abab">
               <l>Awake to the sound of my <rhyme
type="dvpp:rhymeFeminine" label="a">sighing</rhyme>, </l>
               <l>Breeze, gentle breeze of the night ! </l>
               <l>Like an echo awake thou, replying </l>
               <l>To my grief's bitter wailing aright.</l>
            </lg>
```

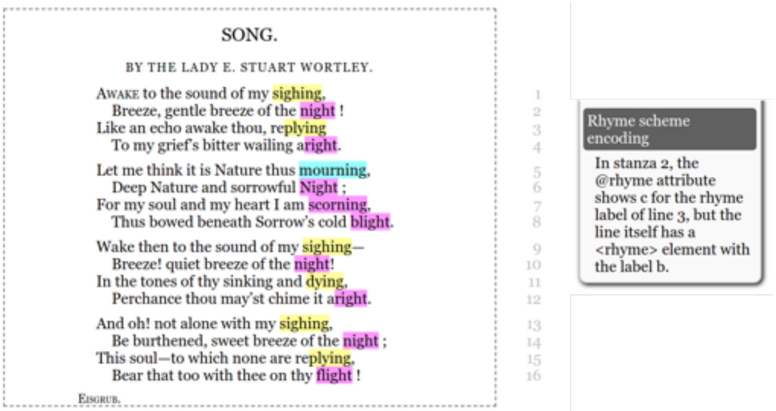
Example 1. An Oxygen code template for tagging rhyme.

```
  <rhyme type="dvpp:rhyme${xpath_eval(if (matches('${selection}', '[aeiouy]
[^aeiouy]+[aeiouy]') and not(matches('${selection}', '^[^aeiouy]*[aeiouy]
[^aeiouy]+ed?$'))) then 'Feminine' else 'Masculine')}" label="${ask('What label?',
generic, 'a')}">${selection}</rhyme>${caret}
```

5. Rhyme-labelling tools

8 The rhyme-labelling protocols in our project are a little unusual. First, we use both the `@rhyme` attribute on `<lg>` to specify the rhyme pattern for a stanza, and also the `@label` attribute on individual `<rhyme>` elements to re-iterate those labels. This duplication has two functions: first, it is easier to analyze the rhyme-scheme of a stanza while its encoding is relatively uncluttered, so we do this before adding the `<rhyme>` elements to individual lines. Secondly, the duplication acts as a check on the accuracy of the line-level labelling. If an encoder assigns `@rhyme="cbcb"` to the stanza, but then inadvertently gives the third line `@label="b"`, our diagnostic rendering (of which more below) will catch this and report the error (figure 8).

Figure 8. Diagnostic report catching rhyme label inconsistency.



SONG.

BY THE LADY E. STUART WORTLEY.

Awake to the sound of my sighing,
  Breeze, gentle breeze of the night !
Like an echo awake thou, replying
  To my grief's bitter wailing aright.

Let me think it is Nature thus mourning,
  Deep Nature and sorrowful Night ;
For my soul and my heart I am scorning,
  Thus bowed beneath Sorrow's cold blight.

Wake then to the sound of my sighing—
  Breeze! quiet breeze of the night!
In the tones of thy sinking and dying,
  Perchance thou may'st chime it aright.

And oh! not alone with my sighing,
  Be burthened, sweet breeze of the night ;
This soul—to which none are replying,
  Bear that too with thee on thy flight !

Eisgrub.

Rhyme scheme encoding

In stanza 2, the @rhyme attribute shows c for the rhyme label of line 3, but the line itself has a <rhyme> element with the label b.

9 The second respect in which our rhyme-labelling differs from many conventional approaches is that we do not restart our rhyme labelling from `"a"` with every new stanza, as is exemplified for example in the TEI Guidelines chapter on "Verse" (TEI 2019). Instead, we maintain a consistent use of the same label for the same sound throughout a poem, no matter how long it is. If "love" is labelled as `"a"` in stanza one, then "dove" will be labelled `"a"` in stanza 57. The reason for this is that we are interested in the effect of rhyme patterns across stanzas as well as within them. It is certainly arguable that there is no plausible effect on the reader of the repetition of a sound dozens of stanzas apart, but it is actually more difficult to specify an arbitrary distance beyond which rhyme recurrence is not relevant, and use such a determination in our encoding, than it is to follow a single sequence throughout the poem.[2]

10 However, this can make rhyme-labelling very difficult in a long poem; it is hard to remember, when encountering "night" at line 300, that "sight" and "flight" were actually rhymed 200 lines earlier and have the label `"b"`. Therefore we have devised a tool based on an XSLT transformation that assists in this process. The encoder specifies the rhyme they are about to tag (say "night"), and the following process is run:

- Every instance of the word "night" tagged as a rhyme throughout the entire collection is collected. This might collect (for example) 40 instances of "night."
- For each of those instances, all the rhymes tagged in the same poem with the same label as "night" are retrieved. This might collect 250 words rhyming with "night."
- For each of those rhymes, all instances of those forms tagged as rhymes throughout the collection are retrieved. This might garner 300 words plausibly rhyming with "night."
- For each of those rhymes, every form tagged in the same poem with the same label is retrieved. This might yield a total of (say) 500 forms which probably rhyme with "night." Of these, there might be 300 distinct forms (because of course there will be many duplicates).
- Finally, each of the distinct values in the list is checked against previously-tagged rhymes in the poem the encoder is working on. If any match, a report is generated as shown in example 2.

Example 2. Report of previous lines in the same poem which are potential rhymes for the selected text.

```
POTENTIAL RHYMES FOR night FOUND IN POEM 478600 night (label = b) a | right (label
 = b) Night (label = b;) blight (label = b)
                 flight (label = b;)
```

11 The encoder is thus able to see that earlier in their poem, “night,” “aright,” “Night,” “blight” and “flight” have all been tagged with `@label="b"`, so that is the appropriate label to use. Since our current collection of tagged rhymes is now well over 30,000, this tool really facilitates the encoding of longer poems.

12 Sometimes, however, a previous instance of a rhyme may be missed, and instead of being given `@label="b"`, a rhyme may be assigned a new label by mistake. When this is noticed, another transformation can be used to fix the problem. The user runs the transformation, and supplies the label which was erroneously assigned, along with the correct label; the transformation then re-labels the entire poem appropriately.

6. CSS styling with rendition/@selector

13 Throughout the nineteenth century, periodicals published poetry in a wide variety of formats and for many different reasons. Sometimes poems were mere filler, inserted to avoid whitespace at the end of a prose article; in other cases, they form part of a prose article on the theme of poetry; sometimes they serve as embellishments for illustrations; and sometimes they are appear as standalone works on otherwise empty pages. Our researchers are interested in connections between form and function, and the relationships of poems to the surrounding material, so it is essential that we capture key aspects of the appearance of poems, including alignment, font styles/variants and size, margins, and indents, and in our HTML rendering of the poems (see below) we try to reproduce this as closely as possible. In our early encoding, we captured these features using inline @style attributes, but the effect of this is to clutter the poem text with complex property-value pairs which make the rest of the encoding harder to work with. As a result, we have

switched to using `<rendition>` elements in the header, and in particular to the relatively rarely-used `@selector` attribute, which enables us to describe the layout of a poem extremely efficiently, especially if it follows a regular pattern. For instance:

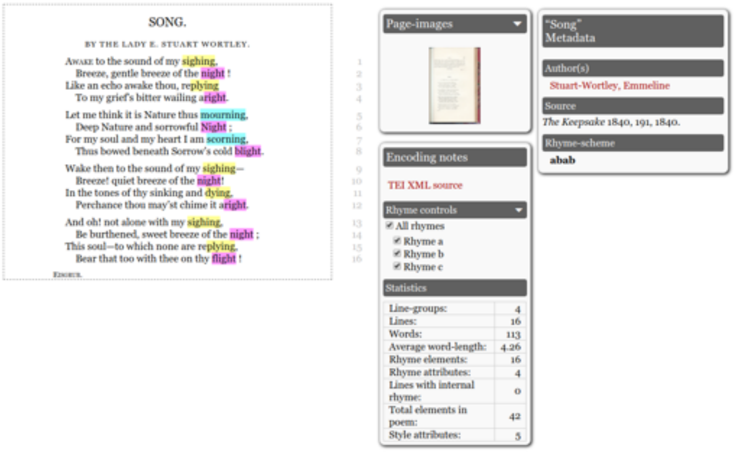
```
<rendition selector="lg">margin: 0.5em auto 0.5em 6em;</rendition>
<rendition selector="l:nth-of-type(even)">margin-left: 1em;</rendition>
```

describes the layout of the example poem used in this article fully, positioning the line-groups within the page, and indenting the second and fourth lines of every stanza. Where initially encoders were spending significant time inserting `@style` attributes in multiple locations, they are now skilled at abstracting the layout of a poem into a couple of `<rendition>` elements in a few seconds.

7. Instant feedback and error reporting

14 Timely diagnostic checking of encoding is extremely important for any serious encoding project (see Holmes and Takeda 2019). In this project, we have found significant benefit in providing a detailed rendering and report (figure 9) for encoders to use as they work, accessible through a transformation scenario available from a single button click in the *Oxygen* editor.

Figure 9. Rendered poem view with feedback report.



SONG.

BY THE LADY E. STUART WORTLEY.

Awake to the sound of my sighing,
  Breeze, gentle breeze of the night !
Like an echo awake thou, replying
  To my grief's bitter wailing aright.

Let me think it is Nature thus mourning,
  Deep Nature and sorrowful Night ;
For my soul and my heart I am scorning,
  Thus bowed beneath Sorrow's cold blight.

Wake then to the sound of my sighing—
  Breeze! quiet breeze of the night!
In the tones of thy sinking and dying,
  Perchance thou may'st chime it aright.

And oh! not alone with my sighing,
  Be burthened, sweet breeze of the night ;
This soul—to which none are replying,
  Bear that too with thee on thy flight !

Page-images

Encoding notes

TEI XML source

Rhyme controls

☑ All rhymes
☑ Rhyme a
☑ Rhyme b
☑ Rhyme c

Statistics

Line-groups:	4
Lines:	16
Words:	113
Average word-length:	4.26
Rhyme elements:	16
Rhyme attributes:	4
Lines with internal rhyme:	0
Total elements in poem:	42
Style attributes:	5

"Song" Metadata

Author(s)

Stuart-Wortley, Emmeline

Source

The Keepsake 1840, 191, 1840.

Rhyme-scheme

abab

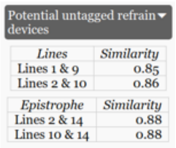
This view allows the encoder to see at a glance the layout of the poem (as they have described it in CSS), check the rhymes (through colour coding and reports on inconsistencies found, as shown in figure 8), as well as providing additional hints for features not yet encoded, as described in the next section.

8. Auto-detection of anaphora, epistrophe and other refrain-like devices

15 In addition to rhyme, we are encoding other features that involve repetitive forms and structures, such as full-scale refrains, anaphora (repeating a word or words at the beginning of a line), and epistrophe (repetition at the end of a line). It is easy for an encoder to miss such features in the

midst of encoding rhyme and layout, so the rendering process includes a component which uses a simple similarity metric to identify potential instances of these features. Figure 10 shows the output of this function as applied to the example poem.

Figure 10. Potential untagged refrain devices identified by the diagnostic rendering.



Potential untagged refrain devices

Lines	*Similarity*
Lines 1 & 9	0.85
Lines 2 & 10	0.86

Epistrophe	*Similarity*
Lines 2 & 14	0.88
Lines 10 & 14	0.88

Here, lines 1 and 9 are similar enough to be tagged ("Awake to the sound of my sighing" / "Wake then to the sound of my sighing"), as are lines 2 and 10 ("Breeze, gentle breeze of the night" / "Breeze! quiet breeze of the night"). Similarly, the phrase "breeze of the night" occurs again in line 14. The automated process enables us to ensure that we catch such echoes even when they are distant from each other in a long poem, and it also gives us a focus, through the (albeit crude) score calculation, to discuss and compare what might or might not constitute a real instance of a refrain device.

9. Conclusion

16 Through a determined focus on developing tools and techniques to improve accuracy and efficiency, we have been able to make a significant improvement in our encoding performance. This has been achieved through continuous dialog between the encoding team and the programmer. We encourage encoders to express their difficulties and concerns, identify awkward and time-consuming parts of the process, and suggest imaginative solutions. We have also been able to leverage the growing dataset of encoding already completed to help with rhyme encoding in particular. Other significant factors include the instant feedback available as a diagnostic rendering of the poem while encoding; a project-level diagnostic process which continuously measures and reports on our progress through the dataset; and the fact that the programmer is also required to do significant amounts of encoding himself, so he encounters first-hand the problems the team is facing, and the frustrations and inadequacies of the solutions provided so far.

APPENDIXES

Appendix 1. Access to code

Detailed code samples have not been included here due to length constraints, but the project codebase is open-source, and can be accessed at https://hcmc.uvic.ca/svn/dvpp/. The following files relate specifically to this presentation:

- The Oxygen project file, which contains all the code templates for keyboard shortcuts and the Schematron QuickFix for auto-tagging a complete poem.
- The XSLT for poem rendering, which includes all the diagnostic feedback code.
- The XSLT for finding rhymes across the whole collection.

BIBLIOGRAPHY

Holmes, Martin, and Joseph Takeda. 2019. “Beyond Validation: Using Programmed Diagnostics to Learn About, Monitor, and Successfully Complete Your DH Project.” In *Digital Scholarship in the Humanities*. Oxford University Press/EADH. 2019. http://dx.doi.org/10.1093/llc/fqz011.

TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.5.0. Last updated January 29. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/3.5.0/doc/tei-p5-doc/en/html/.

NOTES

[1] See example 1. The XPath expression is relatively crude, but it generates correct results almost all the time.

[2] For longer poems, when the 26 letters of the alphabet are exhausted, we restart with `"a1"`, `"b1"`, `"c1"` etc., followed by `"a2"`, `"b2"`, `"c2"` as necessary.

AUTHORS

KAITLYN FRALICK

Masters student in the Department of English at the University of Victoria.

KAILEY FUKUSHIMA

Masters student in the Department of English at the University of Victoria.

MARTIN HOLMES

Programmer/Consultant, University of Victoria Humanities Computing and Media Centre

SARAH KARLSON

PhD student in the Department of English at the University of Victoria.

Adding Word Annotations into TEI Files.

Bertrand Gaiffe *2019-07-05* Introduction

1 Ajouter des annotations dans un fichier XML

1.1 Description du problème

Very often, we face the problem of adding some annotation into an XML file. For instance, we add propernames (name, placeName) into a pre-existing transcrition. The question this paper adresses is to do this annotation automatically. To be more precise, we suppose we have an XML file, containing for instance,

```
<p>The queen Elisabeth the 2<hi rend="sup">nd</hi> celebrates her
93<hi rend="sup">rd</hi> birthday.</p>
```

and a tabular file (were O stand for outside, B-something stands for begin of something and I-something stands for into something) such as :

The	O
queen	O
Elisabeth	B-<name xmlns=”http://www.tei-c.org/ns/1.0”>
the	I-<name>
2	I-<name>
nd	I-<name>
celebrates	O
her	O
93	B-<num xmlns=”http://www.tei-c.org/ns/1.0”>
rd	I-<num>
birthday	O

And we want to produce automatically :

```
<p>The queen <name>Elisabeth the 2<hi rend="exp">nd</hi>
 </name> celebrates her <num>93<hi rend="exp">rd</hi>
 </num> birthday.</p>
```

The problem can be split in two parts :

- alignment of two texts (the pcdata of the XML file on one hand, and the contents of the first column of the tabular file on the other hand
- insertion of the XML tags (, , ,) into the original XML file

An important property we want to insure is that the text of the XML file is not modified, that is, all the game consists in inserting tags into the XML file.

1.2 Inserting tags in an XML file

We can represent an XML file as a sequence of SAX-like events, namely Text Nodes, ElementStart, ElementStop (plus Comment and Procesing Instruction). For instance, our XML fragment would be :

```
  ElementStart("p",0) Text("The queen Elisabeth the 2") ElementStart("hi",0)
Text("nd") ElementStop("hi",0) Text( celebrates her 93) ElementStart("hi", 0)
Text(rd) ElementStop(hi, 0) Text(" birthday.") ElementStop("p", 0)
```

> The second argument of ElementStart and ElementStop, represents the fact the the Element comes from the XML document and not the tabular file.

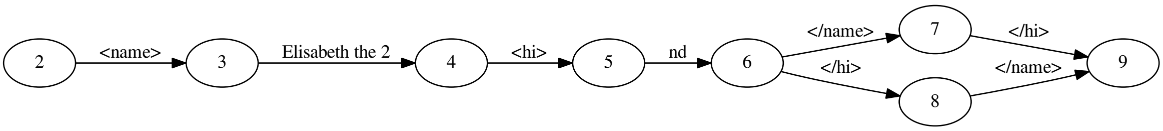
Supposing we know where the elements from the tabular file have to be inserted, we would an equivalent representation except for the fact that we do not know wether

```
  "ElementStop("hi",0)" happens before or after "ElementStop("name", 1")"
```

and wether

```
  "ElementStop("hi", 0")" happens before or after "ElementStop("num", 1)"
```

A good representation of this is a Direct Acyclic Graph. The nodes of the graph are positions in the file and the edges are SAX like events: Now, a simple parsing algorithm (such as Earley's)



using the grammar of the XML language, gives the solution (the upper branch is syntactly incorrect), namely :

```
<p>The queen <name>Elisabeth the 2<hi rend="exp">nd</hi>
 </name> celebrates her <num>93<hi rend="exp">rd</hi>
 </num> birthday.</p>
```

To be a little more precise, the parsing may produce :

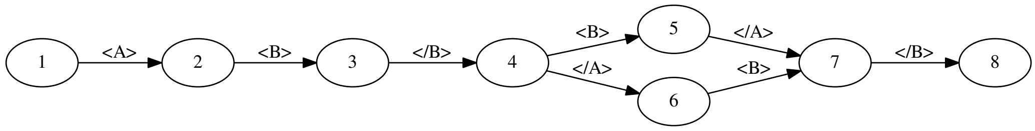
- an error if no solution exists
- a grammar that enumerates the solution(s) if there is at least one *[Note: The grammar generates the language that is the intersection between the DAG and XML language.]* .

1.3 Error management

It may happen that there is no solution to the parsing. This corresponds to situation of the type

```
  <A>...<B>...</A>...</B>
```

The current version of the tool only makes corrections when it encounters the error. This means it closes the "B" and reopens it in the previous example. This leads to the following DAG *[Note: we have the constraint that EStop(B) before EStart(B) and EStop(B) before EStop(A)]* : Now the parsing one the lower branch will succeed.



2 Aligner annotations et fichier XML

In order to build the DAG mentionned in the previous sections, we have to align the text contents of the XML file and the text contents of the tabular file. This is done by a dynamic programming algorithm. The result is an alignment path, that is couples of indices form the XML file and the tabular one respectively.

3 The tool and the way to use it.

The tool is a java jar file. It takes two arguments: the XML file and the tabular one. The contents of the tabular file must match the contents of the XML file (up to small differences such as white spaces, capitals, diacritics, etc.). The output is a XML file (the program outputs one of the productions of the result grammar).

In pratice, however, a TEI file contains more than a text too be annotated, it contains a teiHeader, notes, perhaps variant readings (app), etc. Therefore, usaually, we first fold the parts that should not be annotated, then extract the text, run a tool on that text, transform the resulting tabular file so that it matches our requirements, reinsert the annotations and finally unfold the teiHeader, the notes, the apps, etc.

On an example, it looks like what follows:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>all the header here</teiHeader>
<text>
<body>
<div>
<p>Some text<note>and a note</note> into the body</p>
</div>
</body>
</text>
</TEI>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader xml:id="id1"/>
<text>
<body>
<div>
<p>Some text <note xml:id="id2"/> into the body</p>
</div>
</body>
</text>
</TEI>
```

Some	<w xmlns=”http://www.tei-c.org/ns/1.0” pos=”D”>
text	<w xmlns=”http://www.tei-c.org/ns/1.0” pos=”N”>
into	<w xmlns=”http://www.tei-c.org/ns/1.0” pos=”P”>
the	<w xmlns=”http://www.tei-c.org/ns/1.0” pos=”D”>
body	<w xmlns=”http://www.tei-c.org/ns/1.0” pos=”N”>

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader xml:id="id1"/>
<text>
<body>
```

```
<p>
<w pos="D">Some</w>
<w pos="N">text</w>
<note xml:id="id2"/>
<w pos="P">into</w>
<w pos="D">the</w>
<w pos="N">body</w>
</p>
</div>
</body>
</text>
</TEI>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>all the header here</teiHeader>
<text>
<body>
<div>
<p>
<w pos="D">Some</w>
<w pos="N">text</w>
<note>and a note</note>
<w pos="P">into</w>
<w pos="D">the</w>
<w pos="N">body</w>
</p>
</div>
</body>
</text>
</TEI>
```

4 Conclusion et perspectives

An earlier version of this programme *[Note: this earlier version dit not the alignement of the texts ; it relied on indexes given by the tagger we use, it did however the parsing.]* was used to insert annotations in part of speech and lemmas (tei:w with attributes pos and lemma) into a rather large corpus (5300 texts) of french works (mainly novels). The property of not changing any character in the contents of the files was a strong requirement in this context. Moreover, the syntactic corrections were good hints towards encoding mistakes in the unanotated xml files.

4.1 Running the program(s)

There are actually three programms:

- ”alignXMLAndTab.jar” take an XML file and a tabular file (and an optional column number) and outputs the tabular file with 2 columns added that are indexes into the XML files or the positions of the words (corresponding to the columns number).
- ”faireComp” builds a ”companion file” *[Note: name borrowed from Eric de la Clergerie and is ”passage” project.]* from the output from step one. This program depends on the precise from of the tabular file.
- ”Mixer2” takes an XML file and a companion file and produces the XML file together with the annotations from the companion file.

To the programs, we may add xsl style sheets that fold ans unfold the parts that should not be annotated. And the style sheet which extarcts the text to be given to some external annotation tool that will produce a tabular output.

4.2 Parsing modulo un schéma

So far, the tool does not take a schema as a parameter, therefore, the result is not guaranted to be, for instance, TEI valid. This could be a possible extension that would require:

- taking a schema into account !
- consider two possibilities when facing a syntactic error such as <A>...<B>...</A>.....</B>:
    - The late correction (were the error is encountered) <A>.....<B next=”id1”>.....</B></A><B xml:id=”id1”>......</B>
    - The early one, on the opening element whose closure is encountered at the error point. <A next=”id2”>..</A><B><A xml:id=”id2”>..........</A>......</B>

It might happen, that one of the possibilities is valid and the other not. Therefore, extending the program to take a schema into account would imply coping with the two types of error correction.

The main program and some utilities are available at `https://github.com/bgaiffe/Annotations`

Opportunities and challenges of the TEI for scholarly journals in the Humanities

Anne Baillot, Le Mans Université

The landscape of scholarly journals in the Humanities has been changing rapidly over the past decade. From unaffordable, printed journals conceived as a tool for building up reputation in selected circles, they are now growing into the open access field, implementing various forms of peer review and reflecting on accessibility and circulation. The evolution in mentalities made necessary by changes in the scholarly ecosystem involves an adaptation of the publishing technologies and, more generally, of the management of the publishing process. In this overall context, this paper will concentrate more specifically on strategic issues (accessibility, challenges of openness) and on mechanisms of reputation building (reviewing formats, their implementation and their possible evolutions), both in their relationship to the technical opportunities offered by the TEI.

First, the paper will address a series of publishing issues that are likely to be eased by using the TEI as a format for scholarly journals (among others questions of archiving, accessibility, interoperability, connection to the data). In a second step, it will draw from concrete examples to define the specific challenges involved by the use of the TEI in scholarly journals in the Humanities. The Open Edition Journals platform, the *Journal of the TEI* and the overlay platform episciences will serve as examples and lead to a discussion of the concept of publication itself.

By addressing these questions, the paper offers a reflection on the meaning of the TEI for scholarly communities, and on its capacity to contribute to implementing not simply cosmetic, but truly epistemic changes in the Humanities.

Modeling FRBR entities and their relationships with TEI: A Look at HallerNet bibliographic descriptions

The aim of this paper is to discuss the mapping between FRBR (Functional Requirements for Bibliographic Records) and TEI carried out to create bibliographic records and their relationships. Although some work has been done in this area (Hawkins, 2008), on the differences about database modeling and markup modeling (Eide, 2015) or XML and linked data (Ciotti and Tomasi, 2016), this paper will argue that the TEI Guidelines may be suitable for creating bibliographic records and their relationships.

FRBR (IFLA, 2008) defines four bibliographic entities (Group 1: work, expression, manifestation and item) amongst other entities such as Person, Concept, Event, Place, etc. while the TEI allows to encode "structured bibliographic citation[s]" with <biblStruct> and "a loosely-structured bibliographic citation[s]" with <bibl>. For illustration purposes, this paper will use HallerNet portal devoted to Albrecht von Haller (Bern, 1708-1777), a key figure of the Swiss Enlightenment, and his circle of friends and collaborators. The website comprises both digital editions and a large collection of 35.600 bibliographic records amongst other types of objects. HallerNet has simplified and adapted the FRBR abstract model to define only two bibliographic entities — due to the current state of the bibliographic records — and their relationships: a *work* (a distinct intellectual or artistic creation) and its *manifestations* (the physical embodiment) using the TEI elements <bibl> and <biblStruct> respectively and four potential relationships (*embodimentOf*, *isPartOf*, *isAReviewOf* and *isASuccessorOf*) encoded with <relatedItem>.

Since the TEI Guidelines cover both metadata and data, its vocabulary and syntax can go beyond the representation of texts and facilitate the creation of bibliographic catalogues that group records into "families" based upon some shared characteristics – e.g. same content in different languages or different editions of the same work. Although the abstract model discussed here is a preliminary work that needs to be implemented in the near future, we expect that it will enable further browsing and discovery of records.

Bibliographic references

Ciotti, Fabio and Francesca Tomasi, «Formal Ontologies, Linked Data, and TEI Semantics », *Journal of the Text Encoding Initiative* [Online], Issue 9 | September 2016 - December 2017. URL: http://journals.openedition.org/jtei/1480

Eide, Øyvind, «Ontologies, Data Modeling, and TEI», *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015. http://jtei.revues.org/1191

Hawkins, Kevin S, «FRBR Group 1 Entities and the TEI Guidelines», 2008 TEI Annual Members Meeting, held November 6–8 2008, in London, England, United Kingdom.

IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report. As Amended and Corrected through February 2008* (http://www.ifla.org/VII/s13/frbr/frbr_2008.pdf).

The Prefabricated Website: Who Needs a Server Anyway?

Martin Holmes and Joseph Takeda

ABSTRACT

Project Endings, a collaboration between digital humanists and librarians, is devising principles (https://raw.githubusercontent.com/projectEndings/Endings/master/principles.txt) for building DH projects in ways that ensure that they remain viable, functional, and archivable into the distant future. Endings principles cover five components of project design:

- Data
- Products
- Processing
- Documentation
- Release Management

Previous *Endings* work has focused on Data and Products (Holmes 2017; Arneil & Holmes 2017) and diagnostic tools for monitoring project progress (Holmes & Takeda 2018 and 2019). This presentation will deal with the mechanics of Processing, focusing in particular on building large static sites which are resilient because they have no requirement for server-side technology at all. We will use the *Map of Early Modern London* (MoEML) project as a case study.

INDEX

Keywords: TEI and beyond: interactions, interchange, integrations and interoperability, TEI environments and infrastructures, TEI and publication, TEI and sustainability

1. Introduction

1 *Project Endings*, a collaboration between digital humanists and librarians, is devising principles (https://raw.githubusercontent.com/projectEndings/Endings/master/principles.txt) for building DH projects in ways that ensure that they remain viable, functional, and archivable into the distant future. Endings principles cover five components of project design:

- Data
- Products
- Processing
- Documentation
- Release Management

Previous *Endings* work has focused on Data and Products (Holmes 2017; Arneil & Holmes 2017) and diagnostic tools for monitoring project progress (Holmes & Takeda 2018 and 2019). This presentation will deal with the mechanics of Processing, focusing in particular on building large static sites which are resilient because they have no requirement for server-side technology at all.
2 We will use the *Map of Early Modern London* project (MoEML), one of the flagship Endings project, as a case study. Comprised of 2,000 TEI source files and 15,000 distinct entities, MoEML is a densely interlinked project that requires a sophisticated build process to create its website structure, the historical Agas Map interface, editions of primary source documents, various indexes and

gazetteers, and encyclopedia entries. As a flagship Endings project, MoEML has been a testbed for the scalability of the Endings principles. The MoEML site has 9,000 HTML files, 26,000 XML files, and over 5,000 images, and is around 2GB in size. Our presentation will cover a number of key techniques in the build process, including:

- Validation, validation, validation: XML, HTML, CSS, and TEI `<egXML>` example code is validated at every stage of the build process.
- Diagnostics to check all links and targets.
- Unique query-free URLs for all entities
- Generating the gazetteer, which includes every variant spelling of every placename.
- Pre-generating HTML fragments for AJAX retrieval for every entity.
- Processing and rationalizing `<rendition>` elements and `@style` attributes.
- Using document type taxonomies to build sitemaps and breadcrumb trails.
- Filtering of images to include only those actually used.

2. Why build a static site?

3 When in early 2019 the server which was hosting the tei-c.org website died, the *WordPress*-based main site disappeared from the Internet for a considerable time. Since *WordPress* is a database-dependent system, a single central database host is required to run it, and until a new server could be brought up hosting that database, the site remained unavailable. However, there was no such problem with the TEI Guidelines, which are statically-built and available in multiple locations as a matter of course.[1] A static site, however, can be replicated endlessly. All digital humanities projects will eventually end (Kirschenbaum 2009; Rockwell et al 2014), and their products will transition into minimally-curated archival hosting; static sites have much more chance of survival, availability and replication if they have no server dependencies.

4 The world of large-scale software development is also coming to similar conclusions for slightly different reasons. The JamStack initiative (JamStack.org) is also championing “modern web development architecture based on client-side JavaScript, reusable APIs, and prebuilt Markup,” in the interests of “better performance,” “cheaper, easier scaling,” “higher security,” and a “better developer experience.” Long-term archivability is not a primary goal of JamStack; instead, one of

their motivations is that a static site is far more easily deployed across Content Delivery Networks such as Akamai because it has no reliance on centralized back-end data sources such as a databases. Like JamStack, *Endings* advocates products based on pure client-side HTML5, CSS, and JavaScript.

3. The build process

5 *MoEML*'s static build process, which is managed by *Apache Ant*, takes the densely-encoded, tightly-linked XML collection created by our team and builds from it a massive, loosely-coupled collection of web resources comprising everything we can possibly imagine an end-user might want to see. Before we start, though, we first check whether the current state of the collection is worth building into anything at all. We validate (RELAX NG), we validate again (Schematron), and we check coherence (does every link point to something meaningful?), consistency (does everything conform to the encoding guidelines and the editorial guidelines?), and completeness (does everything mentioned actually exist?) via our diagnostic processes (Holmes & Takeda 2019). If a single file is invalid, or a single link is broken, or an id is used for two different entities, the build fails and the process stops. A website with errors is not worth building.

6 It is worth contrasting this rigorous suite of validation processes with the pre-*Endings* form of the *MoEML* website, which was based on an *eXist* XML database and to which project staff uploaded new and changed materials as they finished them (or thought they had finished them), when it occurred to them, or (sometimes) accidentally, while uploading other materials. Articles were "published" containing links to other articles not yet written, or person records not yet added to the personography. One encoder would add an item to the bibliography with a new id, while another happened to use the same id for a location; both documents would be uploaded, and, at best, links would break and, at worst, the processing would fail to handle the error and break the site. Such issues were not rampant, but they were omnipresent. We will have no more of that.

7 It should also be noted that these sorts of errors were not only caused by encoders; developers of the *MoEML* site had to be very careful that any code committed to the eXist server was error-free. Testing code changes required a parallel hosting environment, which was an additional burden to maintain, and to keep synchronized with the live site. But by checking the validity of our outputs in the static build, we also necessarily ensure that our processes work: if everything in the build is valid, then, at the very least, the code itself can be compiled and it produces valid documents.

Of course, this does not ensure that the code functions precisely the way we want it to, but, as we discuss later, the static build process gives the project time to ensure that the processes work as expected.

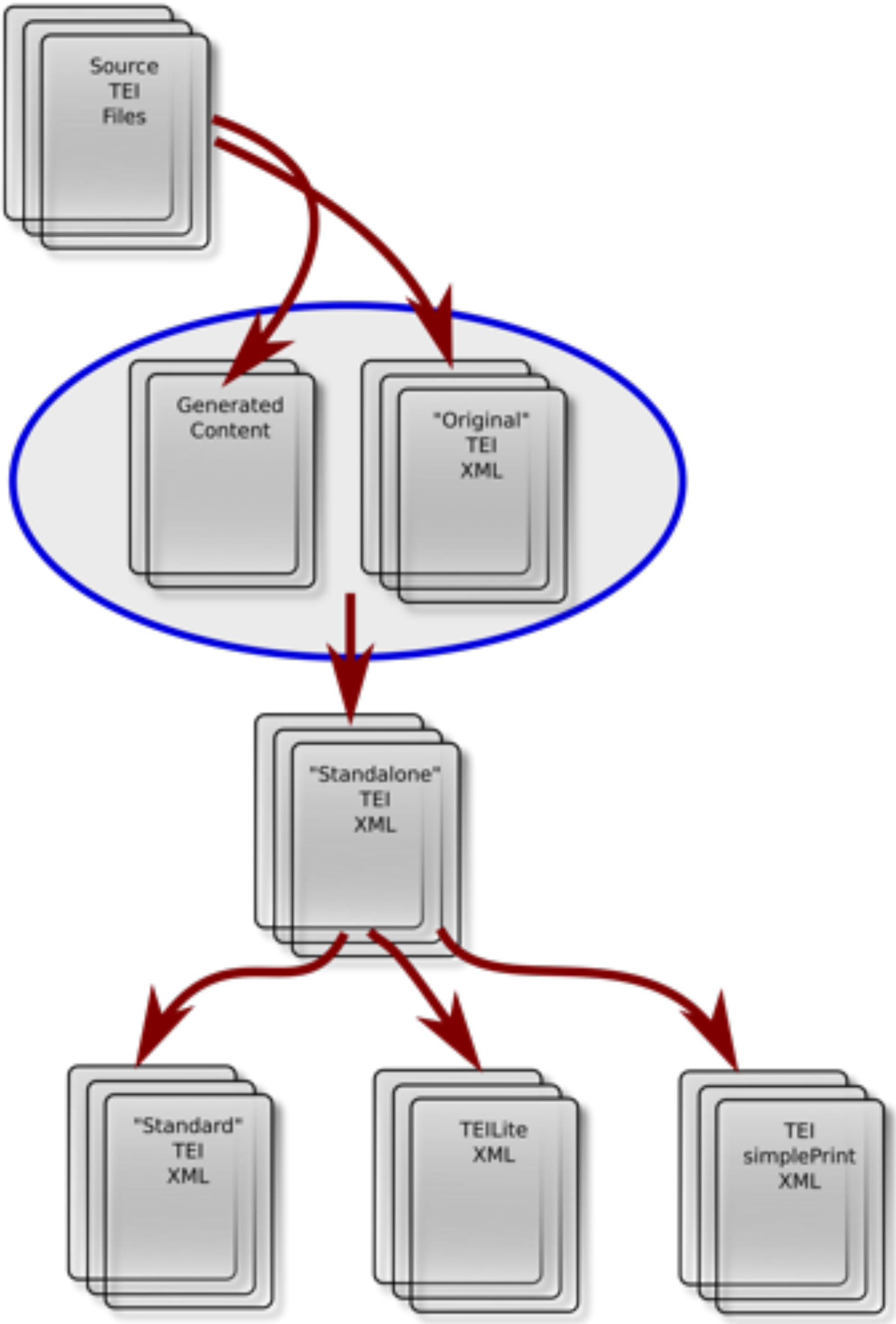
8 Assuming all the validation tests pass, the first stage in building the website is to make more TEI XML. Lots of it, in fact. We build five different versions of our XML collection (see figure 1). Holmes (2017b) provides a full description of the rationale behind this process, but the main justification is that we want to ensure that any future user who comes to our project looking for an XML document can likely find one that is tuned as closely as possible to their needs. We provide XML designed to best represent the praxis of our own project ("original" XML), XML designed to be less esoteric and that aligns more with standard TEI practices ("standard" XML), XML designed to be most amenable to generic external processors (TEI Lite, TEI simplePrint), and XML designed to be detached entirely from the rest of the collection, free of external links and dependencies ("standalone" XML). This is how we end up with 26,000 XML files, from a starting collection of only 2,000. As soon as each new version of the XML is created, you may easily guess what we do with it. We validate it. If any file fails validation, the build stops. This is also when we create a wealth of new files that did not exist before, including the project gazetteer and a range of compiled indexes and similar materials whose information is inherent to the original XML, but which can now be made explicit and tangible. More of this in the next section.

9 Finally, we begin to generate web products. A number of core principles govern the structure and organization of those products:

1. *Every entity* (location, person, bibliography entry, organization, article, etc.) has a *unique id*, and *every unique id gets its own individual page* on the site.
2. *No URL is ever abandoned.* If an id is changed or an entity is removed from the collection, a page is still generated, redirecting to the new version of the id, or to an explanation of what has happened. Linked Open Data requires stable identifiers, so we have a responsibility to maintain them indefinitely.
3. *Every page stands alone and complete in terms of content*; everything referenced in the body of the page (people, places, bibliography items etc.) is included into the page itself, so that if the page becomes detached from its context (if, for example, someone saves a local copy to use while disconnected from the internet), it will continue to work. This of course means that there is massive duplication of data across the site, but we don't care. The entire site still comes in smaller than an HD movie.
4. *All pages live together in the same folder.* This makes for a very large folder, but it means that linking is trivial and reliable, and URLs are easily remembered and typed.

Finally, after the website content is generated, all its pages and associated CSS files are validated with the W3C's VNU validator. As always, any invalid file causes the build to fail.

Figure 1. Generation of multiple TEI output formats.



4. Advantages: you can build anything

10 A major advantage of building the entire site offline is that we can run processes across the entire dataset to build any resource we like, no matter how time- or cycle-consuming it may be. A simple example is the “A-Z Index,” which lists all 9,000 `@xml:id`s used in the project and provides information about the entity to which each id refers. This is an essential resource for *MoEML* encoders, who are often creating new globally unique `@xml:id`s for entities in the project; having a list of all `@xml:id`s not only prevents duplicates ids, but also ensures that encoders can check whether or not the person or place that they are creating already exists in the project. Just-in-time creation of the index is not a feasible option: the server-side construction of the page, which is nearly 10MB, would be slow, even on a powerful server. But, by creating this page ahead of time, the page downloads and renders reasonably rapidly. We also produce a plain-text list of all the ids for faster access and searching by encoders who may be on a slow internet connection.

11 Similarly, it would be impractical to generate the gazetteer of early modern London, which aggregates and groups the thousands of variant placenames across the project, from the source data on a live server. Before implementing the static build, this resource was manually compiled in a semi-automated process. Now we create these documents with the rest of the project, which ensures that these documents not only reflect the current state of the data, but are also completely valid HTML before they are published.

12 The offline build also allows us to take advantage of multi-step processes that would be very difficult to manage in a just-in-time rendering scenario. We make great use of the TEI `<rendition>`/`@selector` mechanism, which uses CSS selector syntax to specify TEI elements to which rendition descriptions apply when encoding presentational aspects of the input. In our build process, we use a two step process during the creation of the “standard” XML to resolve these CSS selectors. For each document that has a `//rendition[@selector]`, we create a temporary XSLT identity transform that converts the CSS selectors into XPath statements, which are then used as the `@match` value for a sequence of `<xsl:template>` elements. We run that transformation against the source document to create the “standard” version of the XML, adding `@rendition`s that correspond with the predefined `<rendition>` in the header. In our standalone process, we then take all `@style` attributes on elements and abstract them into `@rendition` pointers to `<rendition>`

elements in the header. Then, in our HTML creation, all of the `<rendition>` elements are turned into class selectors in the header of the HTML and, accordingly, all `@rendition` attributes are converted into `@class` values.

5. Disadvantages

13 The primary disadvantage of this approach is of course that it involves deferred gratification. Builds take a long time, and they often fail to complete due to invalidities or other errors. It may be hours before an encoder or author can see the results of their work in the context of the built site, and this is particularly frustrating for those who are encoding primary source documents and trying to capture for reproduction rendering features of the original text.

14 However, patience is a virtue and cultivating it is no bad thing. Instant gratification is not a feature of scholarly discourse; compared with waiting for a journal article to be published, waiting a couple of hours to see the latest draft of your document in all its glory is scarcely a hardship. This virtue also extends to the discipline around the public release of complete new versions of a site. Rather than a "rolling release" publication model, where on any given day, the state of the site is inconsistent, incoherent, and unpredictable, a static build process demands a fixed released process, akin to the model of editions of a print text; each edition (delimited by project-specific milestones) is clearly labelled and identified, and always coherent, consistent, and complete. As we have learned from the TEI's incremental releases of the Guidelines, this is a far superior approach, as such releases are much easier to maintain and archive over the long term.

15 In addition, we do provide shortcuts through the build process for local testing of individual files. Our build can be parameterized by supplying one or two specific document ids as input, and in that case, the entire build runs for only those documents and the results are visible within a minute or so.

6. Conclusion

16 We will conclude by summarizing the intent and principles governing our build process:

- Everything that can be pre-fabricated should be pre-fabricated.
- Everything that could conceivably be useful should be created and included.
- Redundancy is beneficial; in fact it is elegant. If the same personography entry is replicated in fifty pages that mention that person, then good; any of those pages can now be used outside the context of the collection without loss.
- Patience is a virtue: let your build take a long time; let your releases be well-separated.

BIBLIOGRAPHY

Arneil, Stewart, and Martin Holmes. 2017. "Archiving form and function: preserving a 2003 digital project." DPASSH Conference 2017: Digital Preservation for Social Sciences and Humanities, Brighton, UK.

Holmes, Martin. 2017a. "Selecting Technologies for Long-Term Survival." SHARP Conference 2017: Technologies of the Book, Victoria, BC, Canada. https://github.com/projectEndings/Endings/raw/master/presentations/SHARP_2017/mdh_sharp_2017.pdf.

———. 2017b. "Whatever happened to interchange?" *Digital Scholarship in the Humanities*, Volume 32, Issue suppl_1, April 2017, Pages i63–i68. https://doi.org/10.1093/llc/fqw048.

Holmes, Martin, and Joseph Takeda. 2018. "Why do I need four search engines?" Japanese Association for Digital Humanities Conference, Tokyo, Japan. https://conf2018.jadh.org/files/Proceedings_JADH2018.pdf#page=58.

———. 2019. "Beyond Validation: Using Programmed Diagnostics to Learn About, Monitor, and Successfully Complete Your DH Project." In *Digital Scholarship in the Humanities*. Oxford University Press/EADH. http://dx.doi.org/10.1093/llc/fqz011.

JamStack: Modern web development architecture based on client-side JavaScript, reusable APIs, and prebuilt Markup. n.d. https://jamstack.org.

Kirschenbaum, Matthew. 2009. "Done: Finishing Projects in the Digital Humanities." *Digital Humanities Quarterly*, Volume 3, Issue 2. http://digitalhumanities.org:8081/dhq/vol/3/2/000037/000037.html.

Rockwell, Geoffrey, Shawn Day, Joyce Yu, and Maureen Engel. 2014. “Burying Dead Projects: Depositing the Globalization Compendium.” *Digital Humanities Quarterly*, Volume 8, Issue 2. http://digitalhumanities.org:8081/dhq/vol/8/2/000179/000179.html.

NOTES

[1] One of the perspectives Holmes brought to the *Endings* project was his familiarity with the static build process for the TEI Guidelines, for which we are primarily indebted to the late Sebastian Rahtz, a wise and clever man who realized all this a long time ago.

AUTHORS

MARTIN HOLMES

Programmer/Consultant, University of Victoria Humanities Computing and Media Centre

JOSEPH TAKEDA

MA Student, University of British Columbia

Between freedom and formalisation: A hypergraph model for representing the nature of text

Elli Bleeker, Ronald Haentjens Dekker, Bram Buitendijk

KNAW Humanities Cluster, Research and Development group

Keywords: text modeling, hypergraph for text, TAG, overlap, discontinuous text, genetic criticism

Over the past decades, the question of "what text really is" has been addressed by a large number of conferences, workshops, articles, and blog posts. Taken together, these contributions make clear that our understanding of text is — and has been — constantly in flux. The flexible and heterogeneous nature of text is reflected by the TEI: users of the TEI Guidelines can mix modules and elements in order to come to their ideal encoding model. Still, there often is a significant gap between a scholar's conceptual model of a text and the way a computer understands that text. Put differently: how an encoding looks to a human reader differs from the way that information is stored on a computer.

A clear example of this divergence is seen with nonlinear text, a textual phenomenon often present on historical or literary documents in the form of deletions and additions. Nonlinear text is typically expressed with `<del>` and `<add>` elements that may be grouped with a `<subst>`: `<s>A <subst><del>first</del><add>second</add> </subst> attempt.</s>`. From the perspective of a human reader, this is a partially ordered data: the textual content is ordered until the point where the variation occurs; the deleted and the added word represent two simultaneous paths through the character sequence. From an informational perspective however, the XML data model cannot properly represent this partially-orderedness: in text-centric XML all text and markup are typically ordered. If the `<del>` and `<add>` elements are placed in a different order, then, they would not have the same meaning (see Bleeker et al. (2018) and Dekker et al. (2018)).

In this long-paper, we present the theoretical and practical implications of using the TAG data model[1] and the associated syntax TAGML to express textual features (such as overlapping text structures, nonlinear text, or discontinuous text) in a way that corresponds more closely with the scholar's understanding of it. We will give examples from the field of textual genesis studies, which presents some particularly intriguing requirements for text modeling (see i.a. D'Iorio (2010) or Bleeker (2017)). As TAG considers text as a network of often implicit information, TAGML documents are inherently multi-layered and nonlinear, and combine ordered and unordered information. Accordingly, scholars can digitally represent their understanding of a text in a more explicit and formalised manner. This paves the way to an innovative approach to creating, modeling, and processing textual objects.

References

Bleeker, E. (2017), *Mapping Invention in Writing: Digital Scholarly Editing and the Role of the Genetic Editor*, Ph.D thesis; University of Antwerp.

Bleeker, E., Buitendijk, B., Dekker, R. H. and Kulsdom, A. (2018), 'Including xml markup in the automated collation of literary texts', *Proceedings of XML Prague 2018* .
URL: *https://www.persistent-identifier.nl/urn:nbn:nl:ui:17-93a4e07d-5be3-471a-a77e-83966d02b370*

Dekker, R. H., Bleeker, E., Buitendijk, B., Kulsdom, A. and Birnbaum, D. J. (2018), 'Tagml: A markup language of many dimensions', *Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies* **21**.

D'Iorio, P. (2010), 'Qu'est-ce qu'une édition génétique numérique?', *Genesis. Manuscrits–Recherche–Invention* (30), 49–53.

[1] TAG stands for Text-as-Graph and is a hypergraph model for text developed and maintained by the RD group of the KNAW Humanities Cluster. See `https://github.com/HuygensING/TAG` and `https://github.com/HuygensING/TAG/tree/master/TAGML` (last accessed July 27, 2019).

Highlighting Our Examples: encoding XML examples in pedagogical contexts

Author: James Cummings

The TEI Guidelines use the egXML element throughout the prose and reference pages for containing XML examples. However, many TEI users know little about this element, and most don't even realise that it is not even in the usual TEI namespace, but instead in a TEI examples namespace (http://www.tei-c.org/ns/Examples).

Following on from my paper at TEI2018 (in which I proposed more detailed ways that the TEI Guidelines might handle examples more generally), this paper will look at possible improvements to the egXML element, specifically designed for modern pedagogical uses. When creating TEI ODD customisations as local encoding manuals, users sometimes use egXML to show how encoders should mark up particular textual phenomena, similar to the use in the Guidelines themselves. Expanding this element's functionality could benefit not only the TEI Guidelines, but also all those who include snippets of XML markup in encoding manuals, slides, tutorials, exercises, or anything else possibly derived from (or exported to) a TEI source and beyond.

Building on the kind of syntax highlighting we are familiar with in XML editors and code snippets online, this paper examines the need to highlight arbitrary portions of XML stored in an egXML element. Whether encoding existing resources containing highlighting of XML or wanting to render modern born-digital pedagogical materials, the TEI Guidelines currently recommend no specific way to do this.

This paper looks at a number of possible options for enabling the highlighting of egXML markup, including embedding namespaced elements, out-of-line markup, and byte-offset coding. All of these are summarised, with the problems that they each face, not only in processing, but also in providing flexible methods to enable users to express existing or desired output rendition.

Introducing Objectification: when is an <object> a <place>?

Author: James Cummings

The TEI Guidelines recently (as of TEI P5 version 3.5.0, January 2019) added elements for describing objects and encoding the names of objects. These elements include: objectName, object, listObject objectIdentifier, as well as changes to many other elements to loosen their descriptions slightly. This paper will introduce these new elements to TEI users who many not have had the chance to use them yet, as well as introduce potential uses for the encoding of object descriptions in TEI files.

The paper will not, however, merely introduce these elements, but will also look at changes still to be done in the TEI Guidelines to fully support the description of objects. For example, much of the object description content model is taken wholesale from that for describing manuscripts. The objectIdentifier was based on the msIdentifier (with some important changes), and object still has elements like msContents in its content model. There are many changes that are still needed and this paper actively seeks to involve the community in designing these changes.

Changes are not only required to existing content models of elements, but also to clarify the semantics of the distinctions between elements. Using examples of large objects such as the Central Library of the National Autonomous University of Mexico (UNAM) (c.f. https://en.wikipedia.org/wiki/Central_Library_(UNAM)), which has substantial and important mosaics on its external walls, this paper will ask where the border lines are between what one might consider an object and what might instead be encoded as a place. Concluding that the difference lies in the markup intent (as with so much in the TEI), this paper draws a distinction between encoding for object description, and geo-political entities.

Exploring TEI structures to find distinctive features of text types

Susanne Haaf

Abstract

Speakers deal with text types (e.g. newspaper, letter, leaflet) successfully every day: They are able to apply the proper text type in a given context, for a certain communicative purpose, according to specific social constraints. However, extensive linguistic discussions on the factors that substantially constitute text types have not come to an end, yet.

Among the key distinctive features of text types the textual structuring has been regularly counted in though its importance compared to other factors could not be finally resolved. Today, with large TEI corpora at hand carrying information on (logical and layout) text structures, it becomes possible to automatically evaluate the relevance of textual structuring for the differentiation of text types. In addition, TEI structures can be included in the recognition of other features whose distribution depends on certain text structures. Hence, next to other criteria it seems straightforward to take a closer look at TEI structuring for the extraction of distinctive features for text types.

The current paper presents an approach to identify distinctive features of devotional text types. Three examples, where TEI structuring is considered, are discussed, namely (1) intertextuality as indicated by bibliographic references, (2) repetition of words and phrases in certain structural contexts, and (3) the level text structuredness in general. The features evaluated here were mentioned in previous (predominantly not corpus-based) studies on distinctive features of devotional literature and of text types in general.

The study is based on three 17th century corpora: manuals of devotion (4,057,497 tokens), funeral sermons (6,910,357 tokens), and a reference corpus of diverse text types (21,862,811 tokens). Texts are taken from the Deutsches Textarchiv corpus and are all tagged according to the TEI format DTABf. It will be shown to what extent TEI tagging can help to safely extract these features and to achieve their more sophisticated interpretation.

Keywords

text types, TEI structuring, devotional literature, text linguistics, corpus linguistics

Related Publications (selection)

Haaf, S. (2019). Art und Funktion von typographischen Mitteln zur Textgliederung in erbaulichen Textsorten des 17. Jahrhunderts. Automatische Analyse im Korpusvergleich und qualitative Einordnung. In F. Simmler and G. Baeva, eds. *Textgliederungsprinzipien. Ihre Kennzeichnungsformen und Funktionen vom 8. bis 18. Jahrhundert*. Berlin: Weidler (Berliner Sprachwissenschaftliche Studien 34), pp. 383–410.

Haaf, S. (2016). Corpus Analysis based on Structural Phenomena in Texts: Exploiting TEI Encoding for Linguistic Research. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, 23–28 May 2016, Portorož (SI).

Kesselheim, W. (2011). Sprachliche Oberflächen: Musterhinweise. In S. Habscheid, ed., *Textsorten, Handlungsmuster, Oberflächen. Linguistische Typologien der Kommunikation*. Berlin/New York: de Gruyter, pp. 337–366.

Simmler, F. (1996). Teil und Ganzes in Texten. Zum Verhältnis von Textexemplar, Textteilen, Teiltexten, Textauszügen und Makrostrukturen. *Daphnis* 25 (1996), pp. 597–625.

Stein, S. (2003). *Textgliederung. Einheitenbildung im geschriebenen und gesprochenen Deutsch. Theorie und Empirie*. Berlin/New York: de Gruyter (Studia linguistica Germanica 69).

correspSearch v2 – New ways of exploring correspondence

The webservice correspSearch has been developed since 2014 to aggregate correspondence metadata and offers it to the scientific community for research and retrieval. The data is obtained in the TEI-XML-based "Correspondence Metadata Interchange Format" (CMIF) - developed by the TEI Correspondence SIG. A prototype was presented at the TEI Conference in Lyon in 2015.

Since 2017, the web service has been further developed in a project funded by the German Research Foundation (DFG). At the same time, the data quantity increased from around 25,000 to over 52,000 letters - many editions offer letter metadata in CMIF by now. In order to enable even small edition projects to deliver data in CMIF and to simplify the capture of letter metadata from printed editions, the CMIF Creator was developed in 2018 to allow a convenient browser-based input and processing of metadata into CMIF.

Over time, the development of the web service focused on both, the system architecture and the improvement of the search, which now - in accordance with the ongoing development of CMIF - for the first time does capture letter content as well. In addition, several different editions of one letter can be linked to each other or connected to associated archives. To add on that, correspSearch does offer a map-based, geographical search for writing and receiving locations by now. The API interfaces and networking possibilities of correspSearch have also been extended. With csLink, a JavaScript-based widget to present correspondence networks is now available on GitHub as open source software that can be integrated into any digital edition.

The article will present and discuss the further development of the web service, as well as the community’s experiences with the aggregation of metadata. During the presentation, the second version of correspSearch will be released as public beta.

Bibliography

Dumont, Stefan: „correspSearch – Connecting Scholarly Editions of Letters“, Journal of the Text Encoding Initiative Issue 10 (2016), http://journals.openedition.org/jtei/1742

TEI Correspondence SIG (Hrsg.): „Correspondence Metadata Interchange Format (CMIF)“, 2018.2015, https://github.com/TEI-Correspondence-SIG/CMIF

The Semantic Field *Grace* in Early Modern English

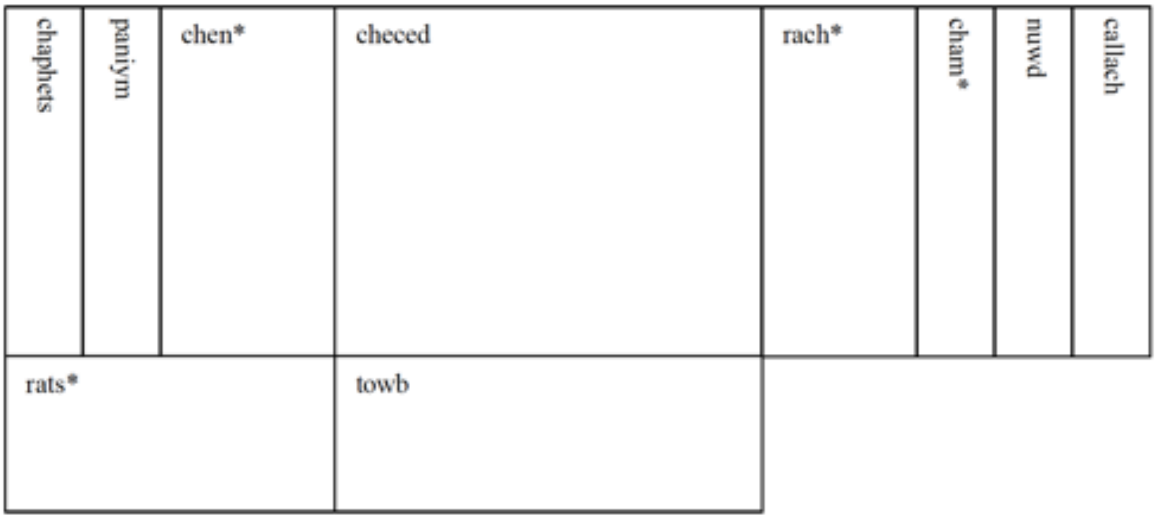
Prof. Brian L. Pytlik Zillig

Dr. Mary K. Bolin

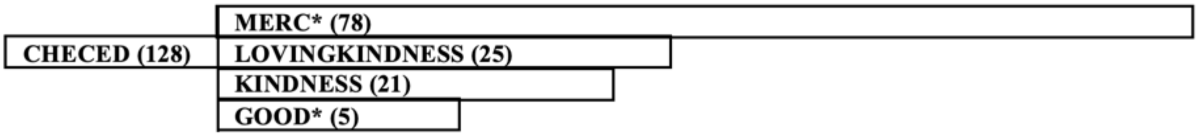
Center for Digital Research in the Humanities
University of Nebraska--Lincoln

The research presented here illustrates the presence of the semantic field *grace* in Early Modern English, using a corpus of well-known and influential texts, and the techniques of Contrastive Linguistics combined with XML tools used to encode and visualize the data. Contrastive Linguistics is a theoretical framework used to compare phonology, morphology, syntax, lexis, and semantics across languages. Examples might include contrasting the vowel phonemes in English and French, or color terms in English and Spanish. Contrastive linguistics can also be used to show contrasts within a single language (or a dialect or historical period within that language). In contrasting semantic areas, the use of semantic fields can map a domain in a way that pictures the domain spatially, showing the relationship of a group of related words. A semantic field is a group of words with related but not identical meanings that all describe or pertain to one domain or semantic area. Once a field is posited, the words can be analyzed and contrasted using a number of methods, including contrastive analysis, componential analysis, semantic primes, and semantic framing. This study uses a semantic field called *grace*, which was originally studied using Bible texts in their original languages, plus English, German, and Latin (Bolin, 1999). The words in the English version of the field are *grace, mercy, compassion, kindness, favor,* and *pity*. This study uses frequencies and standard deviations as the underlying data to visualize the use of the semantic field *grace* in the works of five Early Modern English dramatists: Shakespeare, Jonson, Marlowe, Middleton, and Shirley. Each author's use of the words in the field (including the context) will be contrasted with use by the others, and all will be contrasted with the use of the field in a contemporaneous text, the Book of Psalms from the King James Version (KJV) of the English Bible. Although the Bible was not written in English, the data for this project use only the English of one particular, well-known, and influential translation.

Bolin (1999) mapped the field in Hebrew and Greek in a group of Old and New Testament texts and then mapped the field in English, German, and Latin onto the Hebrew and Greek originals. This crude form of visualization used simple tools that were available 20 years ago. Below is the map of the field in Hebrew, that shows how the words in the field divide up the semantic space



Correspondences in the field were also mapped in other ways. Below is the correspondence of Hebrew *checed* (most commonly translated as *mercy*) with English:



While those simple techniques yielded interesting data, the visualization used in the current research is much more complex, multifaceted, sophisticated, and striking. The data that underlies the visualization includes frequencies and standard deviations, among other statistics. An example of the data is below. It shows the occurrences of the field *grace* (all the words in the field, including variation such as *grace, graceful, mercy, merciful*, etc.) in the texts of the KJV and the five dramatists. The author “various” refers to the KJV text. The frequencies show an interesting difference between the KJV and the Early Modern Drama texts; i.e., that the frequency of the words in the field is higher in the KJV than in the dramatic works. Looking at the words in context is one way to shed light on the reasons for this difference in frequency. The KJV has no standard deviation because only one text is being analyzed, while for the dramatists, the standard deviation shows variation in usage among a number of texts for each author.

Author	Occurs per 10,000	Mean	Std. dev.
Various (KJV)	32.0941	32.0941	0.0000
shirley	15.2739	15.2541	7.9524
middleton	14.8450	14.8463	6.8763
shakespeare	14.8071	14.5976	6.5426
jonson	12.5350	13.0839	6.1316
marlowe	12.2450	12.8340	6.6218

The texts in the corpus were processed with MorphAdorner, a morphosyntactic analysis tool developed by Philip Burns at Northwestern University. This tool identifies part of speech (POS) and lemma information for every word-token in an input text, according to the NUPOS for English schema developed by Martin Mueller. (http://morphadorner.northwestern.edu/morphadorner/documentation/nupos/) Looking at the words in the field in context provides insight and will allow further consideration of their frequencies and distribution. This work uses XML and XML extensions and technologies, including TEI, XSLT, and the XML technology Scalable Vector Graphics (SVG), which is used for the visualization of results.

Works Cited

Bolin, Mary K. (1999). *Grace: a Contrastive Analysis of a Biblical Semantic Field*. Unpublished Master’s Thesis, University of Idaho. https://digitalcommons.unl.edu/libraryscience/6

Archiving a TEI project FAIRly

A. Creamer [https://orcid.org/0000-0002-5286], **G. Lembi** [https://orcid.org/0000-0001-8962], **E. Mylonas** [https://orcid.org/0000-0002-0215], **M. Satlow** [https://orcid.org/0000-0001-7692]
Organization(s): Brown University, United States of America

The Inscriptions of Israel Palestine Project is an online corpus of inscriptions from Israel and Palestine, written in Hebrew, Greek, Latin and Aramaic, dating roughly from the Persian Period to the Arab Conquest. As of spring 2019, it has collected and encoded more than 4000 inscriptions, out of some 10000 relevant texts: we aim to create an exhaustive and easily accessible collection and to enable users to carry out a variety of searches and extensive textual analysis.

The FAIR Principles aim to enhance the ability of machines to automatically find and use digital objects, in addition to supporting their reuse by individuals. The principles are organized under four areas intended to ensure digital objects are *findable*, *accessible*, *interoperable*, *and re-usable.* Following epigraphy.info's mission statement we are applying the FAIR Principles to guide our development of archival formats and processes for our corpus.

As IIP prepared to deposit files in the Brown Digital Repository, we defined formats for ensuring that our files will be as informative, self-documenting and re-usable as possible. Each inscription is contained in a single, XML file, encoded in the well-documented Epidoc subset of the TEI. These files, however, linked to externally maintained controlled vocabularies (using the xi:include feature) and bibliography (using Zotero), in order to facilitate the work of our encoders and ensure consistency. One of our challenges was to incorporate these external data into the robust , stand-alone, archival format.

The archival format of the IIP files is the result of a transformation that writes all the applicable <classifications> directly into the <profileDesc> and also puts complete bibliographic entries derived from Zotero into each file. We will continue to encode using external reference files to ensure consistency, but the archival format should not rely on any external sources.

We will introduce the FAIR Guiding Principles and FAIR Metrics as they apply to epigraphic corpora and TEI encoding, discuss the roadmap for implementation, and look at archival practices beyond FAIR when it comes to preservation of data as well as re-use. While the first steps to making a digital corpus findable and accessible seem straightforward—IIP texts have been ingested into the Brown Digital Repository, have unique and persistent identifiers, rich metadata, and are freely available, we can still improve on both facets. Simple interoperability and re-usability are available through the IIP API in both the production and the archival versions of the corpus, however, it will be important to do further work on controlled vocabularies, shared concepts, and encoding practices in order to enhance both of these facets.

Bibliography

--, 2014. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0 [WWW Document]. FORCE11. URL https://www.force11.org/fairprinciples (accessed 5.10.19).

--, Epigraphy.info [WWW Document], n.d. URL http://epigraphy.info/ (accessed 7.31.19).

Feraudi-Gruénais, F., Grieshaber, F., 2016. Digital Epigraphy am Scheideweg? / Digital Epigraphy at a crossroads? Presented at the Nachnutzung und Nachnutzbarkeit der Forschung im Akademienprogramm Workshop der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste und der Union der deutschen Akademien der Wissenschaften AG „eHumanities“, Düsseldorf. https://doi.org/DOI:10.11588/heidok.00022141

Implementing FAIR Data Principles: The Role of Libraries, 2017. . LIBER. URL https://libereurope.eu/blog/2017/12/08/implementing-fair-data-principles-role-libraries/ (accessed 5.10.19).

Satlow, M., 2002. Inscriptions of Israel/Palestine [WWW Document]. URL https://library.brown.edu/iip/index/ (accessed 7.31.19). https://doi.org/10.26300/pz1d-st89

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Validating selector a regular expression adventure

Author: Syd Bauman

Starting with P3 in 1994 (i.e., over two years before CSS1 was released), the *Guidelines* supported a mechanism to indicate a default rendition, a way of saying ‘all <persName> elements were in italics in the original.’ You would put the name of an element on the gi attribute of a <tagUsage> element in order to indicate which elements had a particular default rendition.

Starting in 2015-10 with P5 2.9.0, TEI introduced a new method for the same purpose (and then phased out the original method). In this new method you specify which elements a default rendition applies to using the Cascading Style Sheets (CSS) selection mechanism — you put a CSS selector on the selector attribute of a <rendition> element. But The TEI only defines <selector> as teidata.text (which boils down to the RELAX NG string datatype).

This struck me as insufficient; formal syntactic validation is in order. Thus I set about writing a regular expression to validate CSS3 selectors. This presentation is about both the process of creating said regular expression, and the result, which is a regular expression just over 18,300 characters long which I believe correctly matches valid CSS3 selectors and correctly fails to match other strings.

Topics to be addressed include the following.

Obstacles to Writing

length and complexity

How do you write such a long expression? The answer is you don’t—you write a program to write the expression. I wrote such a program in Perl, but plan to re-write it in XSLT before the presentation.

confusion

There are some aspects of the CSS3 specification that aren’t entirely clear, at least not to me.

impossible

According to several sources, CSS3 is not regular, and thus it *cannot* be parsed with a regexp. So how was I able to do this? I think there are three contributing factors. I was

Features

output format

The program will generate output in either RelaxNG or XSLT

built-in tests

The output includes a test suite of thousands of CSS3 selectors

case

Because of limitations in RelaxNG’s use of regular expressions, the regular expression produced respects case in some places where it should be ignored.

language

I did not write the portion of the regular expression that tests a BCP 47 language tag, but rather downloaded someone else’s

Resource usage

The regular expression runs very quickly in RelaxNG using jing, and very slowly in XSLT using Saxon.

TEI encoding of correspondence: A community effort

Stefan Dumont, Susanne Haaf, Sabine Seifert

As conveners of the TEI Correspondence SIG and representative for the DTA Base Format, we were repeatedly asked about the TEI encoding of correspondence-specific phenomena. Although the TEI Guidelines contain suggestions, there are still open questions on how to deal with several structural and textual occurrences. This situation led to the idea of holding a workshop to discuss problematic cases of correspondence encoding in TEI, and to develop solutions as well as potential extensions to the TEI.

The workshop was funded by CLARIN-D and was held in October 2018 in Berlin (Germany). We invited early-carrier researchers who deal with TEI encoding and/or correspondence encoding in the course of their daily work, as well as one member of the TEI Council for advice on proposals for TEI extensions. From the participants, we gathered examples of insecurities or problems with applying TEI to correspondence texts beforehand and dealt with these in the workshop.

All problems discussed concern aspects of letter coding: from letter-specific text structures to correspondence metadata. To name but a few, problems with <postscript> or <salute> were treated as well as the use of <correspDesc> for specific correspondence situations. The Correspondence Metadata Interchange Format was also developed further.

Following the workshop, the problems, discussions, and solutions were summarized as handbook-like articles by the workshop participants. The publication “Encoding Correspondence. A handbook on encoding correspondence in TEI-XML and DTABf” will be released step by step as Open Access from July 2019 on with the possibility for the community to review and comment.

In our talk, we want to present this initiative. We will outline the fields of correspondence encoding which were discussed and feature some interesting cases and their solutions. We will also present the course of action applied which to us seemed to evoke fruitful discussions on TEI encoding.

Bibliography

Stefan Dumont, "correspSearch – Connecting Scholarly Editions of Letters", Journal of the Text Encoding Initiative [Online], Issue 10 | 2016, Online since 14 February 2018, connection on 02 April 2018. URL : http://journals.openedition.org/jtei/1742 ; DOI : 10.4000/jtei.1742

Susanne Haaf, Christian Thomas, "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format", Journal of the Text Encoding Initiative [Online], Issue 10 | 2016, Online since 08 August 2017, connection on 27 September 2017. URL: https://journals.openedition.org/jtei/1650; DOI: 10.4000/jtei.1650.

Peter Stadler, Marcel Illetschko, and Sabine Seifert, "Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>", Journal of the Text Encoding Initiative [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016, connection on 02 April 2018. URL : http://journals.openedition.org/jtei/1433 ; DOI : 10.4000/jtei.1433

Long Paper

Keywords

TEI encoding, letters, correspondence

Biographies

Stefan Dumont
Stefan is a research associate at TELOTA, the DH working group at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). He works on interfaces and networking possibilities of digital editions, especially correspondence editions. Since 2017 he has also been coordinator of the project for the further development of the web service “correspSearch” (https://correspsearch.net/). He is co-convener of the TEI Correspondence SIG.

Susanne Haaf
Susanne Haaf is a research associate at the Centre of Language at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and is currently involved in the linguistic projects “ZDL” and “t.evo”. Her work focuses on the constitution and maintenance of historical corpora, TEI encoding of historical texts, issues of text edition, metadata best practices, and the characteristics of historical text types.

Sabine Seifert
Sabine Seifert is a research associate at the Theodor Fontane Archive (www.fontanearchiv.de) at Potsdam University. She contributed to the development of the digital edition “Letters and texts. Intellectual Berlin around 1800” (http://www.berliner-intellektuelle.eu/?en), and has been co-convener of the TEI Special Interest Group Correspondence since 2014. Sabine received her PhD from Humboldt University Berlin with a study on the history of philology in the 19th century.

Advantages and challenges of tokenized TEI

Maarten Janssen

Biography

Maarten Janssen is a researcher at ÚFAL – Charles University, Prague, and the author of the TEITOK corpus platform, a TEI based tool for creating, maintaining, and distributing annotated corpora. TEITOK is used in a growing number of corpora around the world, primarily for historical, spoken, and learner corpora. He is directly involved in a number of TEITOK based corpus projects, including COPLE2, PEAPL, PostScriptum, CORDEREGRA, EFFE-ON, CzeSL, and CoDiaJE.

Keywords: Annotated corpora, Orthographic words, Crossing XML, TEI

Abstract

TEI offers the option to split a text into words/tokens. However, existing tokenized corpora in TEI, such as the TEI version of the BNC corpus, (almost) never really use TEI, but rather are TEI based versions over traditional verticalized texts, with no elaborate TEI markup. When combining full-fledged TEI documents with tokenization, several issues and advantages occur. In this paper, I will discuss the solutions implemented in TEITOK, a TEI based corpus tool, which combines a searchable CWB corpus with editable TEI/XML files.

Proposal

TEI offers the option to split a text into words/tokens. However, existing tokenized corpora in TEI, such as the TEI version of the BNC corpus, (almost) never really use TEI, but rather are TEI based versions of traditional verticalized texts, without using too much of the TEI markup. When combining full-fledged TEI documents with tokenization, several issues and advantages occur. In this paper, I will discuss the solutions implemented in TEITOK, a TEI based corpus tool, which combines a searchable CWB corpus with editable TEI/XML files. Where in a tokenized set-up, a corpus is defined by one word per line, with optionally some semi-XML annotations, in TEITOK a corpus is defined by the tokens inside the XML document itself, defined by the XPath query *//w*. The attributes for each token are then read from the token nodes.

One set of problems that arises from tokenizing full TEI is that elements such as <hi> can break tokens. These problems can be solved by splitting such elements into segments, where the new elements resulting from this can be explicitly marked as repetitions. This is best illustrated with an example: say we have the following TEI fragment: *<hi rend=”underlined”>some under</hi>lining*. Now we would need to tokenize this into two word: *<w>some</w> <w>underlining</w>,* but the <hi> element gets in the way of doing this. Since we cannot use unary tags for <hi>, nor can we break up the tokens, the only available solution is to break up the <hi> while marking the second one as a repetition of the one before to keep the information that there was in fact only one underlining: *<hi rend=”underlined”><w>some</w> </hi><w><hi rend=”underlined” rpt=”1”>under</hi>lining</w>*. Conceptually, this solution is straightforward, although to tokenize automatically can become complicated.

Another set of problems has to do with unary XML elements: the CWB query language is both more powerful, since in CWB all information is related to tokens, meaning searching for tokens is much easier, but relevant information has to be directly connected to tokens. And TEI does not, for instance, explicitly indicate which page a token belongs to since <pb> are unary elements. However, this is easily overcome in a indexed corpus by attaching each token to the last preceding <pb>. In order to establish which page a token appears on, one uses the XPath query ./preceeding:pb[1] to obtain the <pb/> node, where the page number and other relevant information can be obtained.

There are various advantages of tokenized TEI, apart from the primary objective to allow annotating over words. An example is the fact that it voids the need for a @break="no", since word-breaking <lb> are simply those inside a token. And tokens can be associated with several forms, which voids the need for

<ex> and <choice>, since alternative forms such as expanded and regularized forms can be modelled over tokens directly.

An inline tokenized TEI documents with linguistic annotation can be turned into a fully searchable corpus, and the growing number of corpora of TEITOK show the usefulness of this approach. However, in order to be properly searchable, the teiHeader should be treated as structured information and not merely machine readable data: in order to be able to only search in documents with a certain type of encoding, that type has to be in a fixed-value field, and not just in a textual description in the encodingDesc.

Five Centuries of History in a Network

Nathan P. Gibson, University of Munich (LMU)

How should historical researchers encode texts that will be used as primary sources for network analysis? Moreover, how can they link together the various layers of this encoding in a way that explicitly shows the encoder's methodology and interpretations? This presentation is intended to demonstrate and solicit feedback on an experimental procedure applicable to projects producing networks from historiographical texts.

The project "Communities of Knowledge: Interreligious Networks of Scholars in Ibn Abi Usaybiʿa's *History of the Physicians*" (https://usaybia.net)[1] is conducting a network analysis of interactions among Muslim, Christian, and Jewish scholars across five centuries of Abbasid rule in the Near East (750–1258). The primary source text for this endeavor is an Arabic biographical dictionary written by Ibn Abi Usaybiʿa, a physician active in Cairo and Damascus in the 13th century. This *History of Physicians* details teacher-student relations, workplace exchanges, correspondence, oral reports, and other interactions among the scholars mentioned.

The network analysis based on this historiographical text requires a way to proceed from encoding a source text in TEI to producing a dataset for network analysis. Along the way, it must document at least two layers of interpretation: (1) the primary source author's portrayal of events and (2) the researcher's own understanding of these assertions. One way to make these interpretive layers explicit is to isolate individual "factoids" as analytical units. Encoding factoids in TEI and RDF has been recently pioneered by Daniel L. Schwartz for the Syriaca.org project "SPEAR: Syriac Persons, Events, and Relationships." The current project extends this by producing factoids directly from a primary source and integrating them into the encoded text, as well as by further refining the process of curating resulting factoids into nodes and edges for network analysis.

[1] Funded by the German Federal Ministry of Education and Research.

Katharina Prager, Vanessa Hannesschlaeger, Ingo Boerner

Karl Kraus contra ..., or: text contra action

In the project to be presented, the legal papers of the Austrian satirist Karl Kraus (1874-1936) are being edited according to the TEI Guidelines and will be provided digitally and contextualized with Kraus' oeuvre as a whole. Kraus welcomed the reform of the Austrian Press Law of 1922, which marked the beginning of the writer's growing fondness for litigation. In the same year, Oskar Samek became his lawyer. In the course of the following 15 years, they were involved in over 200 court actions together.

The material documenting these actions is the focus of our project. Even though the material's volume (approx. 8000 pages) is a challenge in itself, the most demanding aspect of these documents is their heterogeneity: typescripts, manuscripts, pre-printed forms, carbon copies, and receipts are only some examples of material types we are working with. In addition to the diverse materialities, the heterogeneous functions of the materials (statements, summons, verdicts, correspondences, etc.) pose a challenge as the exact functions of document types have to be understood before the document's qualities can be encoded.

In this paper, we will focus on the document characteristics that are not per se inherent in the text these documents carry, i.e. the documents' functions in relation to real-world processes such as court actions and daily procedures in a lawyer's office. As suggested by Hannesschläger and Andorfer (2019: 8), "the *Text* Encoding Initiative's guidelines, while the unquestionably best approach for encoding text inherent phenomena, reach their limits when used for encoding 'real world phenomena' related to text genesis".

One of the approaches to tackle this problem is to develop a taxonomy in SKOS format to model these processes, i.e. a reusable, TEI-external classification scheme of text types that include different types of juridical documents, court actions, and the procedures in a lawyer's office. In this paper, we will introduce the project, explain our approach and describe the integration of our SKOS taxonomy into the TEI documents containing the texts of our edition by making use of the versatile @ana attribute and the possibilities to include external metadata within the <xenoData> element.

References

Vanessa Hannesschläger, Peter Andorfer. I Want it All, I Want it Now. Literature researcher meets programmer. In Steven Krauwer, Darja Fišer (Eds.). Twin Talks at DHN2019: Understanding Collaboration in DH. Proceedings. Copenhagen 2019. URL: https://cst.dk/DHN2019Pro/TwinTalksWorkshopProceedings.pdf

Modelling linguistic knowledge in TEI: the case of the Vienna Corpus of Arabic Varieties (VICAV)

Karlheinz Moerth (Karlheinz.Moerth@oeaw.ac.at)
Daniel Schopper (daniel.schopper@oeaw.ac.at)
Austrian Centre for Digital Humanities, Austrian Academy of Sciences

VICAV's main objective has been to collect and make available digital material concerning contemporary spoken Arabic varieties, including both linguistically relevant data as well as methodological information with regard to data and tools applied in digitally enabled dialectology. Irrespective of its name, VICAV has been working on a number of quite divergent types of digital language resources such as language profiles, linguistic feature lists, sample texts, bibliographies, dictionaries and documentation of digital tools and workflows. Being situated at the crossroads between diatopic linguistic approaches and research-driven text technology, the project has been serving quite diverse aims: teaching spoken varieties of modern Arabic, teaching comparatistic Arabic linguistics, teaching text encoding by means of TEI as well as experimenting with new technologies.

VICAV was conceived as a 'research lab' allowing to work on new tools and methodological aspects concerning data creation and visualisation. One of the results of the project is an easily deployable and maintainable environment which in its most recent version makes entirely use of X-technologies, data being stored and retrieved via REST directly from a BaseX database, implemented in XQuery, XSLT and XPath. The current interface is characterised by a dual approach to data representation, allowing data to be accessed both through interactive maps and traditional query interfaces. Results are visualised in specialised viewers enabling researchers and students to study the data by juxtaposing and thus comparing them.

One of the challenges of the project was the integration of the quite heterogeneous materials into a harmonised system allowing for flexible extensions without too much overhead for data curators. The system is entirely based on TEI P5 covering many different types of text, ranging from dictionary entries and linguistically annotated texts and corpora to georeferenced bibliographical records and a subject-specific taxonomy.

Keywords: modelling, corpora, linguistics, interface, XML-databases

Bio Daniel Schopper:

After having studied filmmaking, Daniel Schopper earned a degree in German Language and Literature Studies at the University of Vienna, specializing on diary literature around Arthur Schnitzler. His interest in database-supported literary research and text-based theatre studies introduced him into the area of Digital Humanities.
He is group leader at the Austrian Centre for Digital Humanities coordinating the working group on Data, Resources, and Standards, being responsible for the creation and curation of standards-based digital resources like dictionaries or digital editions, covering the full cycle of scholarly data production.

Bio Karlheinz Mörth:
Proceeding from a background in Near Eastern studies (with a focus on modern languages and applied linguistics), Karlheinz Mörth has been working at the interface between modern ICT and humanities studies throughout his academic life. He has conducted research in a wide range of text technological fields taking a special interest in eLexicography, text lexicography, methodologies for the build-up and maintenance of digital corpora, annotation research, and corpus-related encoding standards. Since early 2015 he has been serving as director of the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences, the foundation of which he had helped to prepare in the years before.

Dr. Torsten Roeder

Genesis and Variance: From Letter to Literature

ABSTRACT: The paper examines the nature of textual genesis and textual variance, based on a letter which was later elaborated into an epistolary novel.

KEYWORDS: textual variance, textual genesis, alignment, critical apparatus

The study of textual genesis, as introduced by Karl Lachmann[1] and pursued by many others, makes comprehensible that text is a dynamic medium. During its composition in temporal succession, text is constantly revised, reorganized, and reshaped. Within this process, even context and genre can be subject to change: drafts turn into stories, and whole novels derive from tiny notes.

Documents are what is left of such textual dynamics. Understood as physical manifestations of text, documents appear as stable entities, which can relatively easy be digitized, transcribed, and described. The current standard guidelines for text encoding seem ideal for both machine- and human-readable representations of documents, and they strongly support document-related studies of all disciplines.

However, in order to pursue the question of what "text really" might be, the encoding philologist should also take a look beyond the document and refocus the attention on the examination of textual processes.[2] Which phenomena of dynamics can occur in text, and which encoding stategies are necessary to represent them? How could variance be described and classified, e. g. by distinguishing formal, stylistic, paratextual and contentwise variants?[3] Are the current standards sufficient to represent textual dynamicity, or do they favor document oriented, but less dynamic concepts of text?

The idea of this paper is based on an autobiographical text by Friedrich Rochlitz, an author of the early 19th century, who wrote an eye-witness report of the Battle of Leipzig (1813) to a close friend in Dresden.[4] During the siege, the letter could not be

[1] Cfr. Plachta, B. (2013). *Editionswissenschaft*. 3rd ed. Stuttgart: Reclam, 27–45.

[2] Cfr. Zanetti, S. (2012). *Schreiben als Kulturtechnik*. Berlin: Suhrkamp, 10–16.

[3] Roeder, T. (2018). *Vom authentischen Brief zur durchgestalteten Literatur*. [Blog] Digital Humanities am DHIP. Available at: https://dhdhi.hypotheses.org/5429 [30 Nov. 2018].

[4] Roeder, T. (2018). *Tage der Gefahr*. [Blog] Digital Humanities am DHIP. Available at: https://dhdhi.hypotheses.org/4493 [3 Oct. 2018].

dispatched. Rochlitz continued writing his letter until the war was over, and finally, his friend received a very long and impressively detailed letter.[5] Without the author's consent, the letter was forwarded to other interested readers. Possibly due to positive feedback from the latter, Rochlitz decided to elaborate the letter into a novel, which was published first in 1816, and as a revised edition in 1822.[6]

While the variants between the two printed editions can be described and presented by means of a classical TEI inline apparatus,[7] the textual difference between Rochlitz' original letter and the first edition in print requires a standoff based alignment method, including ambiguities and uncertainties. But does it make sense to combine different encoding methods for one textual history? Does it make more sense to separate variants from witnesses? And which concept of textual dynamics is implied by the decision for a specific markup method?

Outgoing from recent research, the paper presents markup approaches, analysis tools[8] and presentation methods[9] to examine the risen questions on textual genesis and variance, in order to broaden the understanding of textual dynamics for encoding editors and end users as well.

AUTHOR INFORMATION: Torsten Roeder graduated in musicology and Italian studies at the Humboldt University in Berlin. He works in the area of digital humanities since 2007, starting his career at the Berlin-Brandenburg Academy of Sciences and Humanities. His PhD thesis, presented at the University of Würzburg, is based on a TEI edition of 19th century music criticism. Currently he works as Digital Humanities officer at Leopoldina (German National Academy, Halle an der Saale).
Contact: torsten.roeder@leopoldina.org

[5] SLUB Dresden, Mscr.Dresd.h.37,Verm.2°, Saxonica/Sammlungen zur Zeitgeschichte, Tagebuch von Rochlitz über die Begebenheiten in Leipzig (1813).

[6] Rochlitz, F. (1816). Tage der Gefahr. In: *Neue Erzählungen*, 2, Leipzig/Züllichau: Darnmannsche Buchhandlung, 149–365; Rochlitz, F. (1822). Tage der Gefahr. In: *Auswahl des Besten aus Friedrich Rochlitz' sämmtlichen Schriften*, 6, Züllichau: Darnmannsche Buchhandlung, 185–312.

[7] Cfr. TEI: Critical Apparatus. [online] *P5 Guidelines for Electronic Text Encoding and Interchange Guidelines*. Available at: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html [16 July 2019].

[8] E.g. "TEICat: TEI Critical Apparatus Toolbox", http://teicat.huma-num.fr; and "LERA: Locate, Explore, Retrace and Apprehend complex text variants", https://lera.uzi.uni-halle.de.

[9] A preliminal version which compares the two printed editions is available at: http://gefahr.elitepiraten.de/ (password and username: "teiconf"); the original letter will be included until the conference.

Inscriptions, Hieroglyphs, Linguistics... and Beyond! The Corpus of Classic Mayan as an Ontological Information Resource

Franziska Diehr, Sven Gronemeyer,
Uwe Sikora, Christian Prager, Maximilian Brodhun,
Elisabeth Wagner, Katja Diederichs, Nikolai Grube

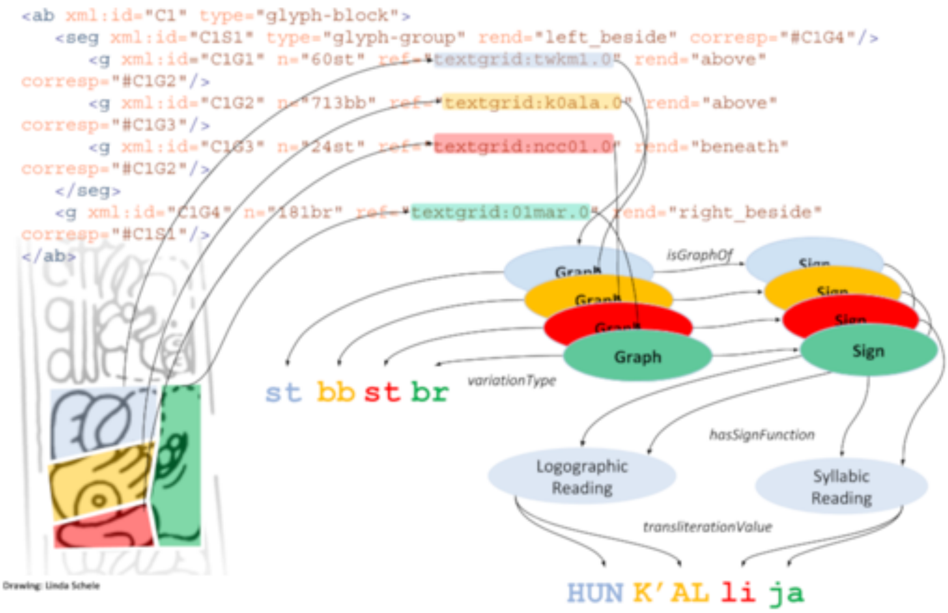
Answering the question "What is text, really?" may be impossible, for 'text' being a most complex resource, fulfilling numerous purposes, manifested in diverse documents types with unique characteristics. To study 'text' using digital methods, some kind of representation is required. The project 'Text Database and Dictionary of Classic Mayan'[1] compiles the hieroglyphic texts written by the ancient Maya in a machine-readable corpus. We do so with an approach fitting the idea of "TEI and beyond": 'Text' is represented as separate information resources, each described by an ontological model representing the specific semantics and complexities of the material. Using different formats (RDF, XML) and standards (CIDOC-CRM, TEI-P5), the inscriptions are encoded in a multi-level corpus:[2] (1) a TEI-all conform schema defining values and rules for the encoding of the text's topological and structural features, (2) a 'Sign Catalogue' for the classification of Maya hieroglyphs[1], and (3) the tool 'ALMAH' for linguistic analyses[2, 5-7].
Maya writing is not yet fully deciphered, not all signs are known, and we still deal with competing reading hypotheses and a missing Unicode character set.[3] To represent the script, we use stand-off markup to enable an interlinked structure between distributed sources: The TEI encoding serves as central data source, embedding other information (Fig. 1). Maya glyphs are grouped in blocks, each usually containing more than one in different arrangements.[4]

[1] For more information on the project, see: `http://mayadictionary.de/`

[2] The corpus further consists of (4) an ontological-based RDF-schema for historical and scholarly information and physical features of text carriers, and (5) the 'Maya Image Archive' for photographic and archival material, for which we use the DARIAH service 'ConedaKOR': `https://classicmayan.kor.de.dariah.eu/`.

[3] There are efforts in this direction [3], but in their current form they do not meet the classification requirements of the Maya script. These challenges are also present in other ancient, non-alphabetic writing systems [4]. The interdisciplinary working group EnCoWS (Encoding Complex Writing Systems) was set up in 2015 for the purpose of harmonising encodings.

[4] Depending on space requirements and aesthetics, individual signs merge, overlap, be infixed, or rotated, depending on the sign shape and space within the block.



```
<ab xml:id="C1" type="glyph-block">
   <seg xml:id="C1S1" type="glyph-group" rend="left_beside" corresp="#C1G4"/>
      <g xml:id="C1G1" n="60st" ref="textgrid:twkm1.0" rend="above"
corresp="#C1G2"/>
      <g xml:id="C1G2" n="713bb" ref="textgrid:k0ala.0" rend="above"
corresp="#C1G3"/>
      <g xml:id="C1G3" n="24st" ref="textgrid:ncc01.0" rend="beneath"
corresp="#C1G2"/>
   </seg>
   <g xml:id="C1G4" n="181br" ref="textgrid:01mar.0" rend="right_beside"
corresp="#C1S1"/>
</ab>
```

Figure 1: Within `tei:g` the value of attribute `@ref` refers to the URI of the graph recorded in the Sign Catalogue. Its ontological structure links the graph to its linguistic expression, to which a transliteration value is assigned.

Using `@rend` and `@corresp` represents this structure by describing the position to the neighboring glyph.[5] The project will encode approximately 10,000 texts. [6] To support the workflow, we developed a parser that creates the according TEI/XML structure out of a project-specific sign number transliteration[2, 2-3].

In our approach, 'text' is understood as a multi-level information resource in form of an ontological corpus, offering different views and access points to the material, providing a holistic environment for studying Classic Mayan.

[5] By using `@corresp` to refer to the neighboring glyph, we mimic a numerical transliteration (similar to the 'Leiden Conventions'), but in a more precise way: With support of the TEI semantics and the XML syntax an unambiguous description of the glyph arrangement is provided.

[6] The data will successively be made accessible under a CC BY-4.0 license on our project portal (https://www.classicmayan.org/) which is currently in the stage of conception. Furthermore, the corpus data will also be published in the TextGrid Repository (https://textgridrep.org/), where they can also be accessed by external users via OAI-PMH. The RDF data of the Sign Catalogue will be also retrievable at the portal via a SPARQL endpoint and also at the TG Rep.

References

[1] Franziska Diehr, Sven Gronemeyer, Elisabeth Wagner, Christian Prager, Katja Diederichs, Uwe Sikora, Maximilian Brodhun, and Nikolai Grube. Modelling vagueness: A criteria-based system for the qualitative assessment of reading proposals for the deciphering of Classic Mayan hieroglyphs. In Michael Piotrowski, editor, *Proceedings of the Workshop on Computational Methods in the Humanities 2018*, volume 2314 of *Workshop Proceedings*, pages 33–44, Lausanne, Switzerland, 2019. CEUR.

[2] Nikolai Grube, Christian Prager, Katja Diederichs, Sven Gronemeyer, Antje Grothe, Céline Tamignaux, Elisabeth Wagner, Maximilian Brodhun, and Franziska Diehr. Textdatenbank und Wörterbuch des Klassischen Maya Annual Report for 2017. *Textdatenbank und Wörterbuch des Klassischen Maya*, (Project Report 5), 2018.

[3] Carlos Pallan Gayol and Deborah Anderson. Achieving Machine-Readable Mayan Text via Unicode: Blending "Old World" Script-encoding with Novel Digital Approaches. In Élika Ortega, Glen Worthey, Isabel Galina, and Ernesto Priani, editors, *Digital Humanities 2018 - Puentes-Bridges: Book of Abstracts*, pages 256–261, Mexico City.

[4] Irene Rossi and Annamaria De Santis. *Crossing Experiences in Digital Epigraphy: From Practice to Discipline.* De Gruyter, Berlin, 2019.

Bibliographies of Authors

Franziska Diehr (Prussian Heritage Foundation, Berlin, Germany): Studied Museology and Information Science in Berlin. Master of Arts awarded in 2013 with an ontology-based data model for the description of scientific collections. Research interests: formal representation of knowledge, especially with regard to the challenges of dealing with vague and uncertain information in the humanities.

Dr. Sven Gronemeyer (University of Bonn, Germany & La Trobe University Melbourne, Australia): Studied Ethnology, Prehistoric Archaeology and Egyptology at the University of Bonn. Doctoral dissertation at La Trobe University on the orthographic conventions of Maya writing and phonemic reconstruction of Classic Mayan. Research interests: Maya epigraphy, Maya history and politics, comparative grammatology, historical and quantitative linguistics.

Uwe Sikora (State and University Library Göttingen, Germany): Studied Egyptology, Classical Archaeology and Akkadian Studies at the Georg August University in Göttingen (Master of Arts). Since 2014 working on different digital humanities research projects with focus on data modeling and data analysis. Research interests: Designing and building information systems.

Dr. Christian Prager (University of Bonn, Germany): Studied Ethnology, Prehistoric Archaeology, and Classical Archaeology at the University of Bonn. Doctoral dissertation about Classic Maya conceptions of their gods. Project coordinator of âĂđText Database and Dictionary of Classic MayanâĂIJ of the North Rhine-Westphalian Academy of Sciences, Humanities, and the Arts since 2014. Research interests: Maya epigraphy, digital epigraphy, comparative studies of writing.

Maximilian Brodhun (State and University Library Göttingen, Germany): Studied Applied Information Science at the Georg August University in Göttingen. Previously conducted Bachelor studies in Applied Information Science at the Georg August University in Göttingen in the department of Health Informatics. Research interests: Graph Databases.

Elisabeth Wagner (University of Bonn, Germany): Studied Anthropology of the Americas, Prehistoric and Protohistoric Archaeology, Ethnology, and Classical Archaeology at the University of Marburg, the University of Cologne, and the Free University of Berlin. Doctoral candidate with a dissertation on the sculptural program of a funerary temple in Copan. Research interests: epigraphy and iconography, sign systems, building programs, craft techniques.

Katja Diederichs (University of Bonn, Germany): Studied German Studies and General Linguistics at the University of Bonn, later completed BachelorâĂŹs degree in Information Processing, Linguistics, and Phonetics at the University of Cologne. Research interests: linguistic information processing, text mining methods, modelling machine-readable metadata.

Prof. Dr. Dr. h.c. Nikolai Grube (University of Bonn, Germany): Studied Anthropology, Ethnology, Assyriology, and Indology in Hamburg. Doctoral dissertation about the development of the Maya script. Postdoctoral qualification on oral traditions of the Cruzoob Maya, thereafter holder of the Linda Schele Chair at the University of Texas, Austin until 2004. Work on Maya writing, iconography, political anthropology, oral traditions, indigenous resistance movements.

Referencing an editorial ontology from the TEI: An attempt to overcome informal typologies

Jakub Šimek (Heidelberg University Library)

The introduction of TEI P5 in 2007 was accompanied by efforts of mapping contents of TEI documents to high level conceptual models like CIDOC CRM. They focused on prosopographical information connecting the textual content with index metadata. Moreover, a flexible use of the `<taxonomy>` element was implemented, allowing for ontology-like thesauri which can be referred to by pointers from an edition.

While these mechanisms for named entities and terms enable powerful indexing, little attention so far has been given to formalizing the ways of dealing with editorial and documentary typologies which are used in attributes like e.g. `@type`, `@function` and `@reason`. These typologies refer to document types, textual and editorial phenomena, the processes of text production and text redaction and similar categories of concepts which characterize the text itself rather than external entities referred to by the textual content.

Attributes like `@type` do not permit the use of pointers to formal conceptual definitions as their expected data type is `teidata.enumerated`, not URI pointers (although the `<equiv>` element in ODD specifications could map different XML components to formal URIs externally).

This paper presents the attempt made at the Heidelberg University Library to enable in TEI documents pointers to definitions of editorial phenomena administrated in a OWL ontology ("heiEDITIONS Concepts") in order to replace `teidata.enumerated` attributes with URI pointer mechanisms. This strategy makes use of a few TEI attributes like `@ana` whose data type is `teidata.pointer` and some additional pointer attributes provided by a schema extension. A "private URI scheme" stated in the TEI header allows the use of abbreviated URI forms.

The goal of this institutional strategy is not only a standardization of the TEI encoding adopted by in-house edition projects and cooperative endeavours but also a transparency in documentary and editorial terminology used in TEI code.

Keywords:

scholarly edition, ontology, pointer, URI, ODD

Recreating history through events

Events are a feature of life in space and time and hence are frequently encountered in witnesses of the human culture.

Coming from diverse backgrounds within the TEI based realm of scholarly editing, the authors identified a common practice in digital editions for encoding events: dateable events are commonly recorded and marked up using e.g. simple date when="*" structures. To date, there is no easy single solution to model events that include basic other event features, mainly referring back to the questions of *what* happened *when* and with *whom*/*what* as subjects and objects of any given event.

There are numerous examples from generically and disciplinarily varying editions where a unified way of describing events could provide added value. Our showcase examples feature material as diverse as We will propose a common strategy on how to encode events in a way that can be easily parsed and extracted from TEI source files.

After careful consideration, we propose minor changes to the <event> and <listEvent> elements. To allow for multiple levels of reporting on events, we will discuss a nested and typed way of distinguishing between describing and described events.

Merging data on historical events from various sources will provide for a closer linking of text editions. It will also provide data for an enriched historical background by linking events across editions. We hope to enable the TEI community to generalize what has lately been successfully adopted for correspondences under the correspSearch label – including a web service –, while staying in line with the generic norms and the specific needs that arise from other types of sources.

Authors: Christiane Fritze, Helmut W. Klug, Stephan Kurz, Christoph Steindl

TEI XML and Delta Format Interchangeability

Authors
Nicholas Cole, University of Oxford; David De Roure, University of Oxford; Pip Willcox, The National Archives

Abstract
This paper will interrogate the close link between TEI and XML, and examine whether the information encoded by TEI could be more easily processed (for certain applications) by expressing it in alternative formats. In particular, this paper will examine the rise of various socalled 'Delta' formats in the JavaScript ecosystem, that are particularly popular with projects developing multi-user, collaborative editors, and which express the structure and features of text by separating the text itself from the list of attributes associated with each block, line, or character. This offers several advantages for processing, and allows such documents to take advantage of an emerging ecosystem.

As this paper examines, for certain applications, such as the representation of negotiated texts, it offers an ability to represent and manipulate rich-text, annotated data in a way that is not possible in traditional TEI formats because algorithms to manipulate delta formats exist that have no comparable (implemented) analogues for XML documents.

This paper will lay out the current state of competing 'delta' formats, the advantages they offer in separating text from the attributes that describe it, the problem that some of these popular but poorly-defined formats would have encounter in trying to express a TEI document in a lossless fashion, whether a translation layer between TEI XML and such a format would be possible to achieve, and whether the academic, text-processing ecosystem would benefit from an alternative format that draws on such ideas, particularly as a transient format (with import/export to/from XML) for editing and data-processing within particular applications. Given the poorly defined state of some other useful but evolving formats, this paper will examine whether efforts by the DH community to standardize versions for academic purposes and link them to the current TEI standard would bear fruit.

Keywords
JavaScript, XML, collaborative editing, annotation, processing

Biographies
Nicholas Cole, Senior Research Fellow at Pembroke College, University of Oxford, studies the political thought of the eighteenth and nineteenth centuries and the history of democratic institutions. He runs the Quill Project on Negotiated Texts which studies the creation of constitutions, treaties and legislation. The Quill software platform (developed with colleagues at the Oxford e-Research Centre) presents a recreation of the original context within which decisions about these texts were made.

David De Roure is Professor of e-Research at the University of Oxford, and a Turing Fellow at The Alan Turing Institute. With a background in hypertext, Web Science and computational methods, David focuses on developing new digital methods in the humanities through collaborations with libraries, archives, and museums. His projects include digital editions, digital musicology, creative computing, innovation in knowledge infrastructure, and the adoption of new technologies.

Pip Willcox is Head of Research at The National Archives. An editor and book historian by background, she has worked on TEI-based projects including the Digging into Data funded

Mapping Manuscript Migrations, which links, analyses and visualizes disparate datasets of manuscript provenance. At The National Archives she leads a multidisciplinary team of researchers and research enablers, working with the historical record of the last 1,000 years. Pip serves on the Board of Directors of the TEI-C.

Case Study TEI Customization: A Restricted TEI Format for Edition Open Access (EOA)

Samuel Gfrörer and Klaus Thoden

Edition Open Access (EOA)[1] is a publishing platform for scholarly monographs and edited volumes that provides authors and publishers with the means to distribute their publications in a variety of formats, e.g. online as HTML, downloadable as PDF and EPUB, or as a printed book. In the new version of the platform, which is currently under development, TEI-XML is used as the central file format. Using the ODD format we are able to specify exactly which parts of TEI-XML we want to use. We can then use the existing TEI infrastructure to create documentation and schemata for our customized TEI format.

For small simple changes creating an ODD customization is simple and feasible: There are some nice tutorials and introductory material and adding or removing some elements or modules is easily being done by adding a few lines into an ODD file. However, because of the subsequent workflow steps, the EOA publishing infrastructure expects TEI documents to have a very specific structure. Using ODD as a central format for describing and enforcing this structure seems to reveal certain shortcomings:

- the semantics of ODD are sometimes slightly vague and hard to figure out from the documentation (How do I know what my ODD defines?)
- some tools in the TEI infrastructure are slightly incomplete or contain bugs
- in ODD, exactly one “content model” for every XML element can be given. TEI provides some general purpose elements (e.g. **`<div>`** or **`<p>`**) which can appear in many different

[1] http://edition-open-access.de, this project is currently funded by the German Federal Ministry of Education and Research, grant number 16OA061.

contexts. In many cases we want to restrict their content in different ways depending on their context, e.g. their position in the XML tree or the value of their **`@type`** attribute (e.g. **`div[@type = 'chapter']`**, **`div[@type = 'section']`**, …). The only way to achieve this with the current state of ODD is to add Schematron[2] rules, which are cumbersome and repetitive to write manually.

In our paper, we give a roadmap of how to customize TEI with ODD considering the specifics of Edition Open Access. We will present a few stategies and tools that address some of the problems mentioned above, e.g. an experimental script to automatically generate an ODD including complex Schematron rules from a Relax NG schema.[3] We hope to initiate a discussion about possible improvements of the ODD format and the TEI infrastructure.

Our roadmap includes the following waypoints:

1. Creating a strict Relax NG Compact schema that covers all the textual phenomena of a scholarly publication in this domain.
2. Learning, understanding and using ODD (TEI modules, the class system, complex restrictions with Schematron).
3. Discovering the problems and limitations of ODD.
4. Finding a solution for EOA that involves the development of tools to generate an ODD file based on the manually create RNC schema in step 1.[4]
5. Discussion: Possible improvements on ODD and TEI.

Authors

SAMUEL GFRÖRER

Samuel Gfrörer studied Computer Science at Freie Universität Berlin. He has been working at the Max Planck Institute for the History of Science (MPIWG) since 2017 as a student worker.

[2] http://schematron.com
[3] https://relaxng.org
[4] Example files and scripts are available at https://github.molgen.mpg.de/EditionOpenAccess/eoa-publication-model.

Since 2019, he is the lead developer of the Edition Open Access, re-designing the web-based platform for Open Access monographs.

KLAUS THODEN

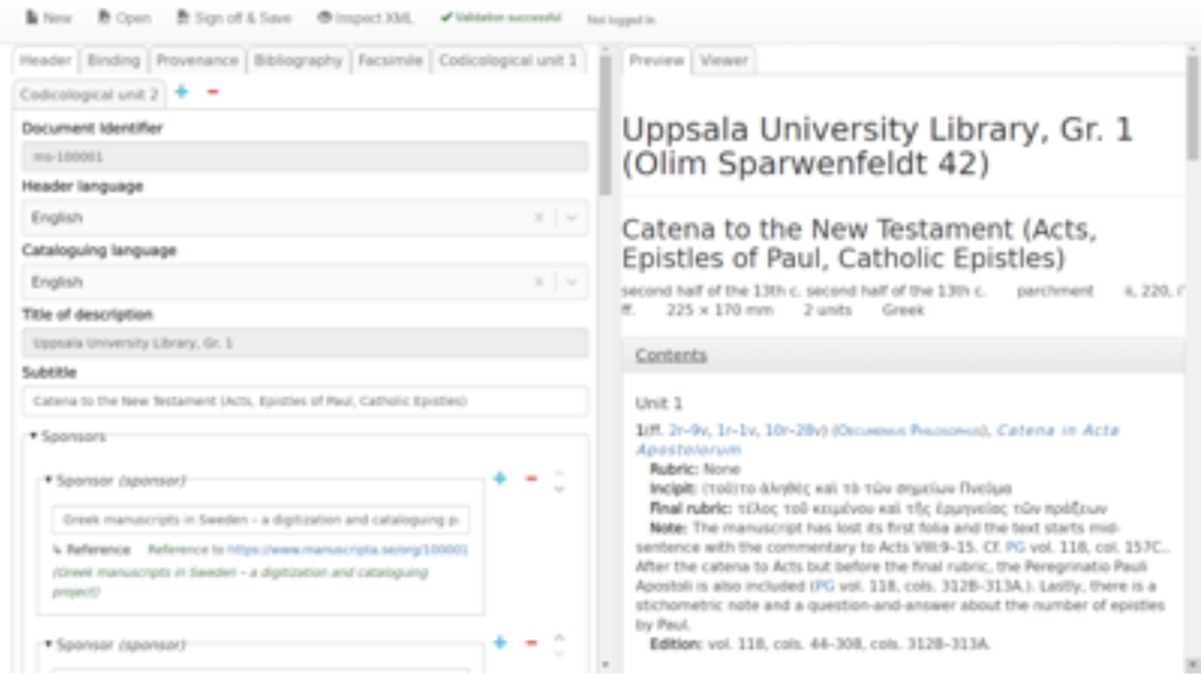
Klaus Thoden is a Research Scholar at the Max Planck Institute for the History of Science in Berlin. He studied German language and linguistics at the Humboldt Universität zu Berlin. He is the Technical Coordinator of Edition Open Access. At the MPIWG, he has been involved in developing an infrastructure for digitizing sources and between 2012 and 2017 has worked in national and international infrastructure projects (DM2E, TextGrid, DARIAH-DE).

Manuscripta – The editor from past to future

Abstract

In this paper we will introduce a web-based editor for TEI-encoded manuscript descriptions in Manuscripta – A Digital Catalogue of Manuscripts in Sweden (https://www.manuscripta.se). Cataloguing is done using an interface which does not require any knowledge of TEI and therefore simplifies and reduces the time required for the cataloguing process. Previously, it has been necessary to use an XML editor which had a steep learning curve and was time-consuming as well as error prone, even with schema validation and detailed cataloguing guidelines.

The interface in the editor is divided into two windows. The left window consists of forms, arranged in tabs according to the structure of the manucript description: Header, Binding, Provenance, Bibliography, Facsimile, and Codicological Unit(s). These tabbed forms are further divided into tabs, in order to minimize the scrolling needed to reach a specific section in the description. The right window consists of two views, a preview and an image viewer, which can be switched via tabs. The preview shows how the description will look when published and the view is updated in real time when the description is changed. The image viewer shows the digitised manuscript, served with via the IIIF API:s using an IIPImage server and Mirador Image viewer. It is also possible to create and edit an IIIF manifest file in the editor. This is done by firstly importing a comma separated file containing filenames, page decriptions, height, and width. The editor then creates a facsimile element where each image has a surface element with graphic and desc elements. It is possible to rearrange both the sequence of the images, and edit the page descriptions. It is also possible to bulk rename the descriptions by page or folio numbers. The facsimile data, together with a selection of data from the manuscript description is then used to create the IIIF manifest.



Document identifier
ms-100001

Header language
English

Cataloguing language
English

Title of description
Uppsala University Library, Gr. 1

Subtitle
Catena to the New Testament (Acts, Epistles of Paul, Catholic Epistles)

Greek manuscripts in Sweden – a digitization and cataloguing p

Reference: Reference to https://www.manuscripta.se/org/100001 (Greek manuscripts in Sweden – a digitization and cataloguing project)

Uppsala University Library, Gr. 1 (Olim Sparwenfeldt 42)

Catena to the New Testament (Acts, Epistles of Paul, Catholic Epistles)

second half of the 13th c. second half of the 13th c. parchment ii, 220, i' ff. 225 × 170 mm 2 units Greek

Contents

Unit 1

1 (ff. 2r–9v, 1r–1v, 10r–28v) (Oecumenius Philosophus), *Catena in Acta Apostolorum*

Rubric: None

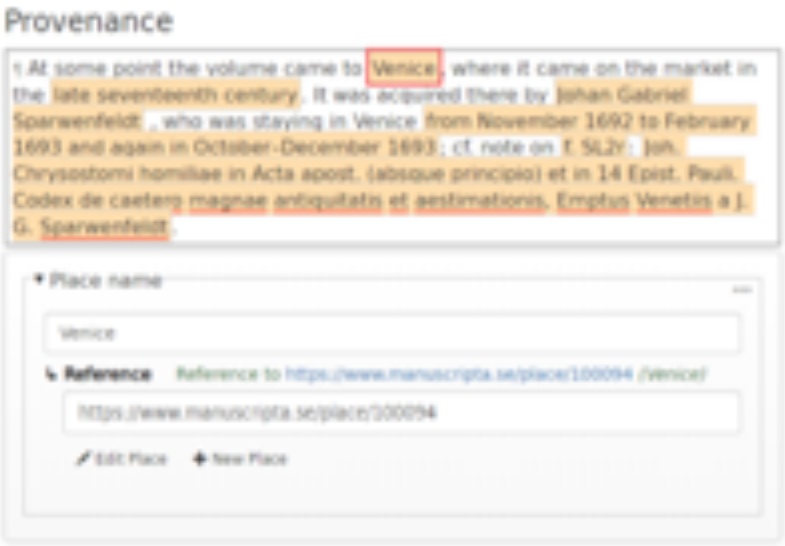
Incipit: (τοῦ)το ἀληθὲς καὶ τὸ τῶν σημείων Πνεῦμα

Final rubric: τέλος τοῦ κειμένου καὶ τῆς ἑρμηνείας τῶν πράξεων

Note: The manuscript has lost its first folia and the text starts mid-sentence with the commentary to Acts VIII:9-15. Cf. PG vol. 118, col. 157C. After the catena to Acts but before the final rubric, the Peregrinatio Pauli Apostoli is also included (PG vol. 118, cols. 312B-313A.). Lastly, there is a stichometric note and a question-and-answer about the number of epistles by Paul.

Edition: vol. 118, cols. 44-308, cols. 312B-313A.

Some parts of the manuscript descriptions consist of running text, e.g. provenance. It is possible to tag words and phrases by selecting the text and then right clicking to get a contextual menu which lists the available elements. The persName, orgName, placeName, and bibl elements can furthermore be linked to authority files, which, in turn, can be edited, or created, in the editor.



Provenance

¶ At some point the volume came to Venice, where it came on the market in the late seventeenth century. It was acquired there by Johan Gabriel Sparwenfeldt , who was staying in Venice from November 1692 to February 1693 and again in October-December 1693; cf. note on f. 5L2r: Joh. Chrysostomi homiliae in Acta apost. (absque principio) et in 14 Epist. Pauli. Codex de caetero magnae antiquitatis et aestimationis. Emptus Venetiis a J. G. Sparwenfeldt.

▾ Place name

Venice

↳ **Reference** Reference to https://www.manuscripta.se/place/100094 (Venice)

https://www.manuscripta.se/place/100094

The schema validation is implemented to be tied to contexts. Each component can have its own context as well as whole documents for each of the different types of documents, e.g. authority documents can have their own contexts. The validation is performed on storing and any validation error is communicated back to the client where you can click on the error description to go to the affected element in the right context.

The Manuscripta-editor is an application package (webapp) on top of eXist-db, an XML database which offers advanced full stack app development with indexing and search functionality, and built-in functions for converting TEI to HTML. In the implementation we use ReactJS, the declarative JavaScript framework, for creating reusable View components. By getting closer to schema-driven design of UI components the cataloguing editor's work is eased if the level of detail is changed in the schema, which would mean configuration rather than programming. With ReactJS' JSX, the XML-ish syntax, it is easy for an editor to combine already generated components if needed for a more complex component. An additional benefit with ReactJS is that there is only a directed one-way dataflow (parent>child) in the components which makes debugging much easier.

While implementations like the TEI Publisher is covering the TEI Processing Model (TEI-PM) with complex text transformations for outputting different media types, navigation, pagination, search, and facsimile display by utilising web components, the Manuscripta-editor also covers other workflow jobs like authority database lookups, advanced templating, editor sign-off, and Schematron rule validation, in addition to schema-based validation. Sharing many goals, like the TEI Publisher, the Manuscripta editor is all about standards, modularity, reusability, and sustainability!

Keywords

medieval manuscripts, cataloguing, eXist-db, ReactJS, reusable components

Native-TEI dialectal dictionary for Bavarian dialects in Austria: data structure, software and workflow

[1,2]Jack Bowers, [1]Philipp Stöckle, [1]Ludwig Maximilian Breuer, [1]Hans Christian Breuer

[1]Austrian Center for Digital Humanities, Austria
[2]Inria - ALMAnaACH, France

This paper discusses the use of TEI in the creation of dually born-digital and print dictionary as part of the Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich 'WBÖ'). Also we discuss the creation of a lexicographic editor tool that allows the non-TEI expert lexicographers to create TEI articles in background of a user-friendly interface.

This work being carried out is a continuation of a legacy project which began in 1913 when data began to be gathered throughout the Bavarian dialect area of the Austrian Empire. The source material being used for the creation of the new articles was collected and elicited using questionnaires and recorded on paper slips. Vocabulary continued to be collected until the 1990's when the analogue records were converted to a TUSTEP database. Recently the database of more than 2.4 million entries has been converted to TEI (Bowers & Stöckle 2018).

At the core of this project are several issues which are particularly significant in the TEI, notably: a) the use of TEI as primary data format for the creation of both a print and digital resource; b) the lexicographic editor tool which provides a user-friendly and open-source alternative to Oxygen XML editor in the creation of systematic and standardized TEI articles using ODD and YAML formatter; c) the structural approach to dialectal dictionary entries in TEI (an under-established/peripheral usage of the module). In our talk we describe the specifics of each of these components of the project and expand upon what has been previously presented about this work in Bowers et al. (2018a,b), focusing particularly on the TEI article structure and the editor tool.

Works cited

Bowers, J., & Stöckle, P. (2018a). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A. U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti, & C. Sporleder (Eds.), *second workshop on Corpus-Based Research in the Humanities (CRH-2)*. Vienna, Austria: Gerastree Proceedings, GTP 1.

Bowers, J., Stöckle, P., Breuer, L. M., & Breuer, H. C. (2018b). TEI born articles on bavarian dialects - updates from the WBÖ. *Digital Humanities Austria 2018*. Presented at the Salzburg. Austria.

An Encoding Strategic Proposal of "Ruby" Texts: Examples from Japanese Texts

Kazuhiro Okada, Satoru Nakamura and Kiyonori Nagasaki

Abstract

We present a novel encoding strategy in "ruby" texts in Japanese. We propose some elements to encode ruby texts, referring to existing ruby encoding in other schemata: **`<ruby>`**, **`<rb>`**, and **`<rt>`**. Our model differs from existing models in fulfilling historical complex ruby attachments.

Keywords: *Japanese texts, ruby encoding, glosses*

In Japanese texts, "ruby" text is a run of text in phonetic scripts attached to a particular portion of the main body of the line (generally Chinese characters) in order to represent the reading of the portion (W3C, 2001, *inter alia*). It is named after the body size name of it in the moveable type printing. It is also called *furigana* as it is allocated to the main texts, which originally developed in East Asian vernacular glossing culture to the Classical Chinese texts.

Ruby today is not, however, a mere annotation to the line: it is more stuck to the particular characters than simple interlinear glosses and, more importantly, it should be read in parallel with the main line, or even as if it were the main line itself. Thus, ruby can be used not only as a simple guide of pronunciation but also as an alternative to the main text. Historically speaking, ruby was also attached to both sides of the main text in order to denote both reading and gloss, especially for borrowing words: even in this case, readers may read any side of ruby in accordance with their preferences. Besides, each side can correspond to different spans of text. These ruby functions are not confined to a simple gloss, which can be encoded within the current TEI framework, and they deserve to be encoded in its own way.

One can find such a two-sided example where a word 打球場, or *billiard hall*, is attached two rubies (figure 1): “ビリヤード” (phonetic transcription of the word *billiard* in Japanese *katakana*) to the whole, and ダキウ (phonetic transcription of the word 打球 in Japanese *katakana*) to the first two characters.

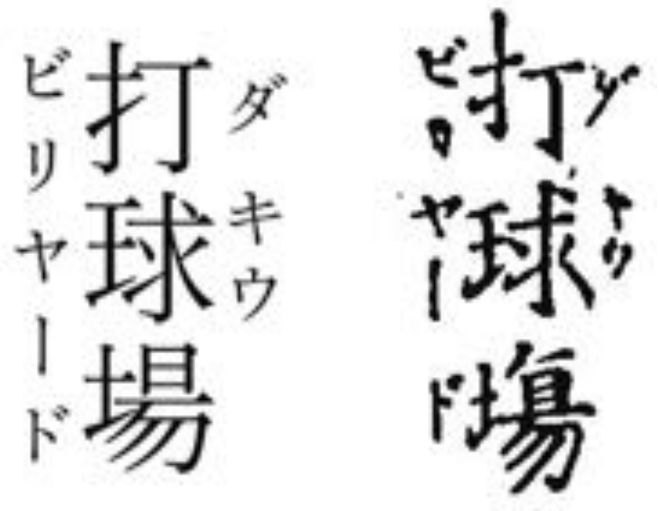


Figure 1. From f. 28v of Niwa (1878) and its modern rendering. Public Domain.

With this case, we will propose to encode this as follows:

```
<ruby>
  <rb>
    <ruby>
      <rb>打球</rb>
      <rt type="p">ダキウ</rt>
    </ruby>
    場</rb>
  <rt type="s">ビリヤード</rt>
</ruby>
```

Example 1. An Encoding Example of the Figure 1.

Here, following existing conventions, **`<ruby>`** element denotes the whole ruby environments; **`<rb>`** element denotes the body texts, where nested **`<ruby>`** elements are allowed inside to encode two-sided rubies; and **`<rt>`** element denotes the ruby texts, where `@`**`type`** describes whether the ruby text in question is primary (`"`**`p`**`"`) or secondary (`"`**`s`**`"`). Primary and secondary rubies are determined according to the text direction: in vertical right-to-left writing, primary rubies are on the right, whereas horizontal left-to-right writing, they will be on the top.

Thus, our encoding strategy encompasses historical varieties. Since ruby has long been one of the major concerns in Japanese texts, its encoding strategy has naturally been proposed from its early days, including W3C's ruby specifications (W3C, 2001; 2013a; 2013b, Hara & Yasunaga, 2002, WhatWG, 2019). Their considerations are quite broad and informative, regretfully limited to modern usages or either confused on the inclusion of interlinear gloss as ruby solely based on layout similarities.[1] Our proposal differentiates interlinear glosses to be encoded by the **`<gloss>`** element, which are attached to a particular place, rather than a particular part of the body (See Estill, 2016 for complexities in encoding of interlinear glosses). This distinguation is of necessary to an encoding to be more proper and simpler.

Bibliography

Hara, S., & Yasunaga, H. (2002). "国文学研究支援のためのSGML/XMSデータシステム: 国文学データ共有のための標準化." 『情報知識学会誌』11 (4): 17–35. doi:10.2964/jsik_KJ00001039453.

JIS (2004). 『日本語文書の組版方法』JIS X 4051. Tokyo: Japanese Industrial Standards Committee.

Kawabata, T. (2014). "HTMLのルビ標準化の現状と課題." 『漢字文献情報処理研究』15: 4–15.

Estill, L. (2016). "Encoding the Edge: Manuscript Marginalia and the TEI." *Digital Literary Studies.* 1 (1). https://journals.psu.edu/dls/article/view/59715.

Niwa, J. (1878). 『竜動新繁昌記 初編』 doi:10.11501/767340.

WhatWG (2019). "Text-Level Semantics." *HTML Living Standard.* 18 July 2019. https://html.spec.whatwg.org/multipage/text-level-semantics.html.

W3C (2001). *Ruby Annotation.* https://www.w3.org/TR/2001/REC-ruby-20010531/.

——— (2012). *Requirements for Japanese Text Layout.* https://www.w3.org/TR/jlreq/.

——— (2013a). "HTML Ruby Markup Extensions." 25 February 2013. http://darobin.github.io/html-

[1]. This is because the existing ruby encodings have focused in realizing the contemporary ruby layouts, especially presented as JIS (2004) or more accessible W3C (2012). See Kawabata (2014) for more references.

ruby/snapshot20130225.html.

——— (2013b). *Use Cases & Exploratory Approaches for Ruby Markup.* https://www.w3.org/TR/ruby-use-cases/.

Authors

KAZUHIRO OKADA

National Institute of Japanese Literature, Japan

SATORU NAKAMURA

The University of Tokyo, Japan

KIYONORI NAGASAKI

International Institute for Digital Humanities, Japan

An Attempt of Dissemination of TEI in a TEI-underdeveloped country: Activities of the SIG EAJ

Satoru Nakamura (The University of Tokyo), Kazuhiro Okada (National Institute of Japanese Literature), Kiyonori Nagasaki (International Institute for Digital Humanities)

One of the missions of Special Interest Group for East Asian/Japanese (hereinafter SIG EAJ)[1] is the dissemination of TEI in Japan. A characteristic of Japan in comparison with TEI-advanced countries in North America and Europe is that the culture which utilizes XML-related technologies such as XSLT and XQuery have not been widespread. In addition, methodologies of research for textuality have not widely been treated among humanities researchers in Japan. In this paper, we discuss how to spread TEI in such TEI-underdeveloped countries based on the activities of SIG EAJ.

One of our activities is the hands-on sessions which handle "Aozora Bunko" texts. Aozora Bunko[2] transcribes public domain works like Project Gutenberg, and about 20,000 texts are available in its original format and HTML. This hands-on session aims to provide not only the opportunities to learn TEI by the converting practice but also the TEI-compliant Japanese texts on the Web. We are also preparing a comprehensive set of tutorials for Japanese resources using GitHub[3]. Sections such as how to write the TEI header, how to encode plays have been created so far. Through these sessions, more than 20 texts were encoded in conformance with Level 3 of Best Practice for TEI in Libraries. In addition, we are developing visualization tools using JavaScript such as CETEIcean[4], and the program which automatically converts Aozora Bunko texts into TEI in the Level 2 of the practice.

By coupling with compiling tutorials and accumulating markup examples targeting Aozora Bunko texts, we have gained in-depth knowledge of markup on modern Japanese texts. Through this activity, the number of people who are familiar with TEI is getting increased. Moreover, in countries where XML-related technologies have not been widespread such as Japan, JavaScript tools can work well to share the benefits of TEI. This point is considered to be helpful in the spread of TEI in other TEI-underdeveloped countries.

[1]. Text Encoding Initiative (2019). TEI: East Asian/Japanese SIG https://tei-c.org/Activities/SIG/EastAsian/.

[2]. 青空文庫（2019). 青空文庫 Aozora Bunko https://www.aozora.gr.jp/.

[3]. TEI: East Asian/Japanese SIG (2019). GitHub repository for TEI: East Asian/Japanese SIG https://github.com/TEI-EAJ.

[4]. TEIC (2019). TEI in HTML5 Custom Elements

https://github.com/TEIC/CETEIcean.

Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings

Tomaž Erjavec and Andrej Pančur

Abstract:

Parliamentary proceedings (PP) are a rich source of data used by e.g. scholars in historiography, sociology, political science, linguistics, and economics and economic history. As opposed to sources of most other language corpora, PP are not subject to copyright or personal privacy protections, and are typically available on-line thus making them ideal for compilation into corpora and open distribution. For these reasons many countries have already produced PP corpora, but each typically in their own encoding, thus limiting their comparability and utilisation in a multilingual setting.

The talk will overview current approaches to encoding PP, with a focus on TEI and TEI-like encodings, on Akoma Ntoso, a standard specifically designed for encoding PP and other legislative documents, and on RDF, also a common approach to encoding PPs. We then motivate and propose a TEI ODD (so, a schema parametrisation and guidelines) for such corpora, based on the TEI module for Transcriptions of Speech. The work on this Parla-CLARIN recommendation started with the "CLARIN ParlaFormat" workshop (cf. https://www.clarin.eu/blog/clarin-parlaformat-workshop) with selected participants who presented their own experiences with encoding parliamentary corpora and gave their comments to the draft proposal by the authors.

These comments have been largely taken into account, and the current Parla-CLARIN recommendation is available at https://github.com/clarin-eric/parla-clarin. The Git repository contains the ODD, the derived HTML guidelines and XML schemas, and example documents. The recommendation presents and discusses the encoding of PP metadata, including the speakers and political parties, the structure of the corpus, the encoding of the speeches and notes, linguistic annotation and multimedia.

The talk concludes with discussing further work, esp. the provision of a set of example documents, the conversion of Akoma Ntoso and RDF encoded PPs into Parla-CLARIN and vice-versa, and other transformation scripts that would operationalise the proposed encoding.

Keywords: Parliamentary corpora, TEI parametrisation, Akoma Ntoso, Linguistic annotation

Author bios:

Tomaž Erjavec is a senior researcher at the Dept. of Knowledge Technologies, Jožef Stefan Institute. His work focuses on developing language resources, esp. as regards their annotation and encoding, both in the fields of language technologies and digital humanities. He is the national coordinator of the Slovenian branch of the CLARIN research infrastructure for language resources and tools and a member of ISO TC37 SC4, .

Andrej Pančur is a research fellow at the Institute of Contemporary History, a national coordination institution for Slovenian branch of DARIAH. Since 2011 he has been working in Research Infrastructure of Slovene Historiography, where he is responsible for technological development, research data, digital editions and the transfer of digital humanities techniques and methods into the Slovenian research area. In 2018 he became the head of the research infrastructure.

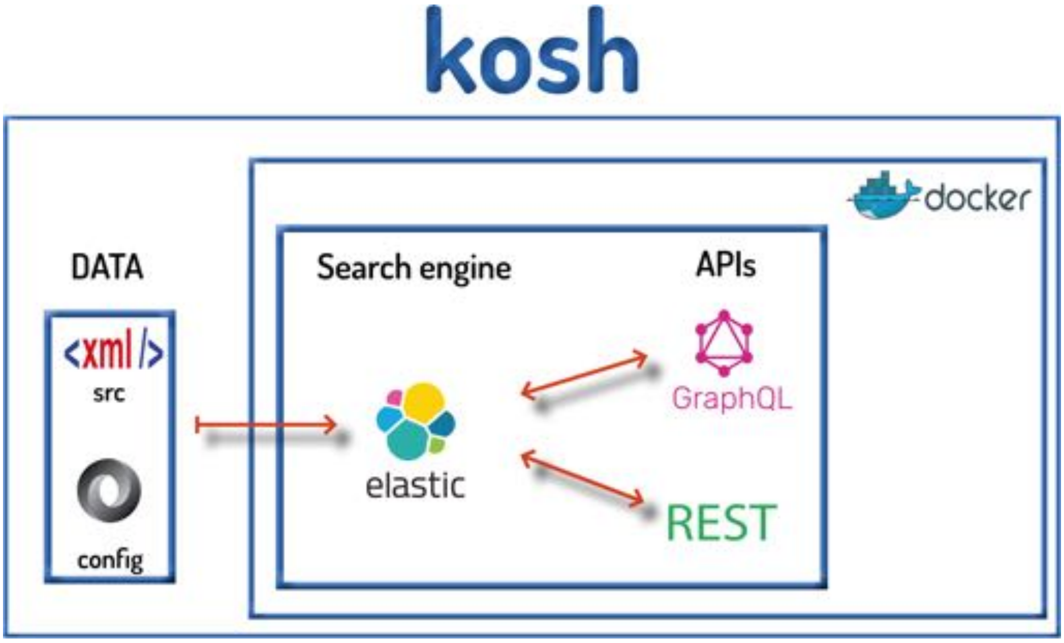
Introducing an Open, Dynamic and Efficient Access for TEI-encoded Dictionaries on the Internet

Francisco Mondaca, Philip Schildkamp, Felix Rau, and Jan Bigalke

1 Most of the TEI-encoded dictionaries in public data repositories are not directly accessible for computational processing. Their use by different applications depends on how each application processes each single dictionary. In the last decades direct computational access to data on the Internet has been provided through application programming interfaces (APIs). APIs provide a centralized access to data and if designed and implemented properly, an efficient access to it. But API development and maintenance requires technical expertise, which can be an obstacle for small and medium dictionary publishers that might not have in-house solutions for this purpose. Against this background we have developed Kosh,[1] an open-source framework, that processes any XML-encoded dictionary and creates two APIs for accessing the underlying lexical data: A REST API (Fielding 2019) and a GraphQL[2] API. The purpose of this presentation is to show how to use Kosh with data publicly available on GitHub, in order to demonstrate how the edition of digitized dictionaries and the compilation of digital-born dictionaries can be supported with an efficient access to the underlying data via APIs.

2 Kosh is an open-source framework developed to access multiple XML-encoded dictionaries. It is generic and flexible, designed to handle dictionaries of different structures and size with a minimal configuration effort.

Figure 1. Simplified Overview of Kosh's Architecture.



3 Kosh processes as input data in XML format that is parsed and indexed into an elasticsearch[3] server. In a JSON (JavaScript Object Notation) configuration file, the paths to the elements to be indexed are defined[4] as also the elasticsearch datatypes of the fields to be indexed e.g., keyword or text. Finally, a Kosh data module requires a dot file (.kosh) containing the index name, the paths to the XML files to be indexed, and the path to the configuration file. With this information, Kosh indexes one or multiple XML files into one index that is accessed by two APIs, a GraphQL and a REST API. If the XML source files are modified, the index is updated automatically. Kosh can be deployed via Docker[5] or natively on Unix systems and a single Kosh instance can provide access to multiple dictionaries.

4 In a GitHub repository, Kosh Data,[6] there are different datasets that show the structure of a data module for Kosh. One of them contains the *Diccionario Geográfico-Histórico de las Indias Occidentales ó América*, a five-volume dictionary compiled by Antonio de Alcedo (De Alcedo 1786, 1787, 1788a,

1788b 1789), which offers a wide description of American toponyms and also, on its fifth volume, a vocabulary with a pioneer approach to descriptive word usage in the Spanish Americas (Lenz 1905–1910:7f). An XML version of this dictionary has been employed in digital gazetteer projects such as HGIS de las Indias[7] and later in Pelagios Commons[8]

5 Based on this XML-encoded version we created a TEI-P5 compliant version. This data can be accessed in two ways: First, through Kosh Data,[9] where modifications to the data can be proposed through pull-requests. Second, via APIs[10] provided by a Kosh instance deployed with a clone of this repository.

6 As data modifications are done at source-file level, and the changes tracked with git,[11] the edition process is open and also reversible. In Kosh the publisher defines which fields should be indexed. For the digitization of printed dictionaries this means that direct computational access can be provided with a coarse-grained encoding. When compiling born-digital dictionaries a few fields can be available at an early compilation stage and as the data gains complexity more fields can be added to the index. This flexible approach to lexical data access allows to unveil datasets that are currently hidden from computer applications and thus users.

BIBLIOGRAPHY

De Alcedo, Antonio. 1786. Diccionario Geográfico-Histórico de las Indias Occidentales ó América. Tomo I. Madrid: Imprenta de Manuel Gonzalez. Available at: https://archive.org/details/diccionariogeogr06alce

———. 1787. Diccionario Geográfico-Histórico de las Indias Occidentales ó América. Tomo II. Madrid: Imprenta de Manuel Gonzalez. Available at: https://archive.org/details/diccionariogeogr07alce

———. 1788a. Diccionario Geográfico-Histórico de las Indias Occidentales ó América. Tomo III. Madrid: Imprenta de Manuel Gonzalez. Available at: https://archive.org/details/diccionariogeogr08alce

———. 1788b. Diccionario Geográfico-Histórico de las Indias Occidentales ó América. Tomo IV. Madrid: Imprenta de Manuel Gonzalez. Available at: https://archive.org/details/diccionariogeogr09alce

———. 1789. Diccionario Geográfico-Histórico de las Indias Occidentales ó América. Tomo V. Madrid: Imprenta de Manuel Gonzalez. Available at: https://archive.org/details/diccionariogeogr10alce

Fielding, Roy Thomas. 2009. Architectural Styles and the Design of Network-based Software Architectures. Irvine: University of California. Available at: https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

Lenz, Rodolfo. 1905–1910. Diccionario Etimolójico de las Voces Chilenas Derivadas de las Lenguas Indíjenas Americanas. Santiago: Imprenta Cervantes. Available at: http://www.ub.uni-koeln.de/cdm/ref/collection/mono20/id/6705

NOTES

1 http://kosh.uni-koeln.de

2 https://graphql.org

3 https://www.elastic.co

4 In XPath 1.0

5 https://hub.docker.com/r/cceh/kosh

6 https://cceh.github.io/kosh_data

7 http://www.hgis-indias.net

8 http://commons.pelagios.org

9 https://github.com/cceh/kosh_data/tree/master/de_alcedo

10 GraphiQL:http://kosh.uni-koeln.de/api/de_alcedo/graphql

Swagger UI (REST):http://kosh.uni-koeln.de/api/de_alcedo/restful

11 https://git-scm.com

ABSTRACT

INDEX

Keywords: api, dictionary, framework, kosh, graphql, rest

AUTHORS

FRANCISCO MONDACA

Cologne Center for eHumanities - University of Cologne

PHILIP SCHILDKAMP

Data Center for the Humanities - University of Cologne

FELIX RAU

Department of Linguistics - University of Cologne

JAN BIGALKE

Cologne Center for eHumanities - University of Cologne

Making Linkable Data from Account Books: Bookkeeping Ontology in the Digital Edition Publishing Cooperative for Historical Accounts

Christopher Pollin and Kathryn Tomasek
TEI2019, University of Graz

Historical accounting records, a genre of manuscript document that can be encoded in TEI/XML, can be a rich source about local social and economic relationships in the past as well for comparative purposes. With the Digital Edition Publication Cooperative for Historical Accounts (DEPCHA), documentary editors and developers in the United States collaborated with digital humanists at the Zentrum für Informationsmodellierung (Centre for Information Modeling) at the University of Graz to apply a bookkeeping ontology to accounts.

The main use cases for this project are the Financial Papers of George Washington (Stertzer), accounting records from a store on the Stagville Plantation in North Carolina (Brumfield and Agbe-Davies), and a day book kept by Laban Morey Wheaton, a businessman who kept a dry goods store in Norton, Massachusetts, between 1828 and 1859 (Tomasek and Bauman).

This paper focuses on a workflow for creating linkable data using the TEI `@ana` attribute. While the Wheaton edition was created using TEI-XML, the data for the Washington Financial Papers and for the Stagville store were created using Drupal and FromthePage, respectively. Transformation to TEI-XML for these records is an interim step for creating the RDF for linking on the Web of Data (Pollin 2019). Examples will be drawn from the Stagville/FromthePage and Wheaton/TEI-XML data. Upcoming work on building a bridge from Drupal to TEI-XML is part of the project's ongoing research agenda.

Part of the workflow for creating RDF includes use of WikiData and OpenRefine. Since the currency, goods, and services exchanged had contextual meanings grounded in both place and time, ongoing work also focuses on mapping information from taxonomies based on the digital editions of the primary sources to Historical Statistics of the United States, a long-standing project in U.S. economic history. Examples will include an alpha-version TEI taxonomy drawn from this source. Future work will involve mapping taxonomies from the Washington Financial Papers onto this TEI taxonomy and expanding it accordingly.

References

Bicentennial Edition: Historical Statistics of the United States, Colonial Times to 1970. https://www.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970.html

Brumfield, Ben, and Anna Agbe-Davies. Encoding Account Books Relating to Slavery in the U.S. South. 2015. https://medea.hypotheses.org/182

DEPCHA prototype: https://gams.uni-graz.at/depcha

Historical Statistics of the United States. Millenial Edition Online. https://hsus.cambridge.org/HSUSWeb/toc/hsusHome.do
Pollin, Christopher. Digital Edition Publishing Cooperative for Historical Accounts and the Bookkeeping Ontology. forthcoming 2019.

Stertzer, Jennifer. Working with the Financial Records of George Washington: Document vs. Data. Digital Studies/Le champ numérique. 2014. DOI: http://doi.org/10.16995/dscn.57

Tomasek, Kathryn, and Syd Bauman. Encoding Financial Records for Historical Research. Journal of the Text Encoding Initiative. 6. 2013. https://journals.openedition.org/jtei/895

Bios

Christopher Pollin holds a Joint Master's Degree in Digital Heritage (EuroMACHS). Currently he is a PhD candidate in Digital Humanities and research assistant at the Centre for Information Modelling (Graz). His work in technical development and data modelling includes the following projects: *STEFAN ZWEIG DIGITAL*, *'Open Access Database Adjective-Adverb Interfaces in Romance'* and *DEPCHA*. The main focus of his work is on *Web of Data* technologies, resource discovery, and web programming.

Kathryn Tomasek, Professor of History at Wheaton College in Massachusetts, is the current Chair of the Board of Directors of the TEI. She was the Principal Investigator for the planning award that DEPCHA received from the National Historic Publications and Records Commission and the Andrew W. Mellon Foundation.

Using Machine Learning for the Automated Classification of Stage Directions in TEI-Encoded Drama Corpora

Authors

Daria Maximova (National Research University Higher School of Economics, Moscow, RU)
Frank Fischer (DARIAH-EU and National Research University Higher School of Economics, Moscow, RU)

Keywords

drama corpora, stage directions, short text classification, machine learning

Abstract

The <stage> tag is a core element for the encoding of drama. The TEI guidelines suggest nine values for its type attribute, which is widely used in large corpora such as the French *Théâtre Classique*, the Shakespeare Folger Library or the Swedish *Dramawebben*. This paper introduces an approach to automatically assign stage-direction types to the TEI-P5-encoded Russian Drama Corpus, RusDraCor (https://dracor.org/). The corpus currently features 144 plays ranging from mid-18th to mid-20th century which makes for 32 753 stage directions with 144,525 tokens.

We selected 18 plays comprising 6,569 stage directions to represent the breadth of the corpus. For the manual annotation we established a clear set of rules to identify the stage-direction types proposed by the TEI guidelines (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-stage.html).

Following the annotation of our subcorpus, we developed a tool for the classification of the remaining plays without human interference. For the conversion of stage directions into feature vectors, we used morphological and semantic data. Our tool in its current state is able to classify different types with an F1 score of approx. 0.75, which means that 3 out of 4 stage directions of any given type are assigned correctly.

Our work will inform a dedicated analysis of stage directions, which after preliminary studies by Sperantov (1998) and Detken (2009) will be based on larger corpora allowing for a description of the evolvement of stage directions over 200 years.

Short Bios

Daria Maximova studies computational linguistics at Higher School of Economics, Moscow.

Frank Fischer is Associate Professor for Digital Humanities at Higher School of Economics, Moscow, and director of DARIAH-EU, the pan-European digital research infrastructure for the Arts and Humanities.

A sign of the times: medieval punctuation, its encoding and its rendition in modern times

Authors: Elisa Cugliana, Gioele Barabucci

Digitally managing punctuation in the editions of medieval manuscripts is one of those issues that initially look like minor details, but later reveal themselves as a tangled web of problems spanning from computer science (how to represent punctuation signs?) to philology (what types of signs do exist?) through epistemology (is the processing of punctuation a mere technical transformation or a valuable part of the scholarship?). The aim of this paper is to address the theoretical aspects of these questions and their practical implications, providing a couple of solutions fitting the paradigms and the technologies of the TEI.

The debate on how to deal with medieval punctuation is a long and still open one. Following Contini (1992), the interpretative edition of a manuscript is the translation of a historically attested system into another system. Accordingly, the philologist should recognize the punctuation system of the manuscript and convert it into a modern one. There are, however, no established universal methods for doing so; most of this work is left to the experience (and taste) of the scholar. In fact, editors often substitute the original punctuation with a modern one. This improves the text readability, but often leads to a loss of textual information.

In this paper we describe how we dealt with the encoding and transformation of the punctuation signs in some German manuscripts of Marco Polo’s travel account. Technically, we implemented a set of general rules (as XSLT templates) and various ad-hoc exceptions (as descriptive instructions in XML attributes). In addition to this, we discuss the philological foundation of this method and, contextually, we address the topic of the transformation of a single original source into different transcriptions: from a ‘hyperdiplomatic’ edition to an interpretative one, going through a spectrum of intermediate levels of normalization. We also reflect on the separation between transcription and analysis, as well as on the role of the editor when the edition is the output of a semi-automated process.

Challenges in encoding parliamentary data: between applause and interjections

Parliamentary data has always been of great interest to researchers in the social sciences and the humanities. There are many initiatives at European and national levels for compiling digital collections of parliamentary data. However, these initiatives use different encoding schemes to present parliamentary data ranging from ad hoc ones, over specific standards for representing legislation (e.g. Akoma Ntoso) to TEI. Akoma Ntoso has been created to make the structural and semantic components of digital parliamentary documents fully accessible to machine-driven processes[1]. However, this encoding standard was not designed to include linguistic annotation. In this regards TEI is more suitable. The Austrian parliamentary record corpus, ParlAT (Wissik & Pirker 2018), which was before only available in a vertical format suitable for analysis in a corpus query system, is now being encoded in TEI. In this contribution, we will present the encountered challenges, often related to the fact, that the Austrian parliamentary records are edited shorthand records of the parliamentary sessions and not transcripts of recordings. The Austrian parliamentary records include a lot of comments within parenthesis and formatted differently than the other text, namely in italics. These comments range from indicating applause or laughter to indicating interjections. We decided to encode these comments in different ways using the elements from the transcription of speech module: to use <incident> for comments indicating applause or laughter and <u> (utterance) for comments indicating interjections[2]. In the contribution, we will discuss the solutions for encoding such comments, regarding the Austrian case but also in relation to a more general scheme proposed by Tomaž Erjavec and Andrej Pančur, the teiParla[3].

References

Wissik, T. and Pirker, H. (2018). ParlAT beta Corpus of Austrian Parliamentary Records. In: D. Fišer, M. Eskevich, F. de Jong, ed., *LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora* In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC2018*. Miyazaki: European Language Resources Association, 20-23.

Keywords

Parliamentary records, TEI, transcript of speech module, encoding applause, encoding interjections

[1] https://unsceb-hlcm.github.io/part1/index-13.html (accessed 11.05.2019)

[2] Parts of this research were supported by a CLARIN Mobility Grant allowing a research visit at the Jožef Stefan Institute as part of CLARIN Slovenia.

[3] https://www.clarin.eu/event/2019/parlaformat-workshop

Using Microsoft Word for preparing XML TEI-compliant digital editions

The paper will introduce the so-called electronic editions prepared in Microsoft Word text processor. This tool was originally developed in 2000 in order to generate so-called vertical format for a corpus manager and in 2008 it was modified in order to output XML TEI format for the Manuscriptorium project (National Library of the Czech Republic 2019). For capturing structural information (headings, indices etc.) and the semantics of the individual parts of the edition (notes, readings etc.) the editors use character and paragraph styles whose application is described in Černá and Lehečka (2016). The methodology also includes rules of transforming the individual styles into XML format according to the TEI P5 Guidelines (The Text Encoding Initiative Consortium 2019). An add-in for Microsoft Word was programmed to help editors to apply formatting and other necessary parts of the edition, e.g. page and line numbers. The author of the paper will focus on DOCX document conversion to XML TEI P5 format. The conversion which consists of approximately 60 sequentially applied XSLT transformations (World Wide Web Consortium 2009) is driven by a specialized application (programmed in C#). In order to keep as much of "the editor's intent" as possible and reduce errors resulting from the transformation process (omitted or duplicated text) an electronic tool was created which extracts plain text (divided into basic text and annotations) from the input (DOCX) and target document (XML TEI P5). This can be carried out by a much smaller number of XSLT transformations (only 6 stylesheets). The output text is subsequently compared via text comparison tools, e.g. WinMerge (2013): differing text chunks point to problematic passages, where transformation to XML format fails. With these tools approximately 240 editions of Old- and Middle-Czech texts (from 1300–1800) was prepared during 10 years, currently attracting new potential users across institutions.

References

Černá, A. and Lehečka, B. (2016). *Metodika přípravy a zpracování elektronických edic starších českých textů. (The Methodology of the preparation and processing of electronic editions of Old Czech texts.)* [pdf] Praha: oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Available at: <http://vokabular.ujc.cas.cz/soubory/nastroje/Methodics/Metodika_pripravy_a_zpracovani_elektronickych_edic_DF12P01OVV028.pdf> [Accessed 14 May 2019].

National Library of the Czech Republic (2019). *Manuscriptorium. Digital Library of Written Cultural Heritage*. [online] Available at: <http://www.manuscriptorium.com> [Accessed 14 May 2019].

The Text Encoding Initiative Consortium (2019). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [online] Available at <http://www.tei-c.org/Guidelines/P5/> [Accessed 14 May 2019].

WinMerge 2.14.0 (2013). Available at <http://winmerge.org> [Accessed 14 May 2019].

World Wide Web Consortium (W3C) (2009). *XSL Transformation (XSLT) Version 2.0.* Available at <https://www.w3.org/TR/2009/PER-xslt20-20090421/> [Accessed 14 May 2018].

Analyzing and Visualizing Uncertain Knowledge: Introducing the PROVIDEDH Open Science Platform

Alejandro Benito (University of Salamanca), **Michelle Doran** (Trinity College Dublin), **Jennifer Edmond** (Trinity College Dublin), **Michał Kozak** (Poznań Supercomputing and Networking Center), **Cezary Mazurek** (Poznań Supercomputing and Networking Center), **Alejandro Rodríguez** (University of Salamanca), **Roberto Therón** (University of Salamanca) and **Eveline Wandl-Vogt** (Austrian Centre for Digital Humanities)

Abstract

Underlying uncertainty in DH research data affects decision-making and persists during the project's lifecycle. This uncertainty will always be present. Thus, efforts in providing technical support for humanistic research should focus on managing and making it more transparent, rather than removing it.

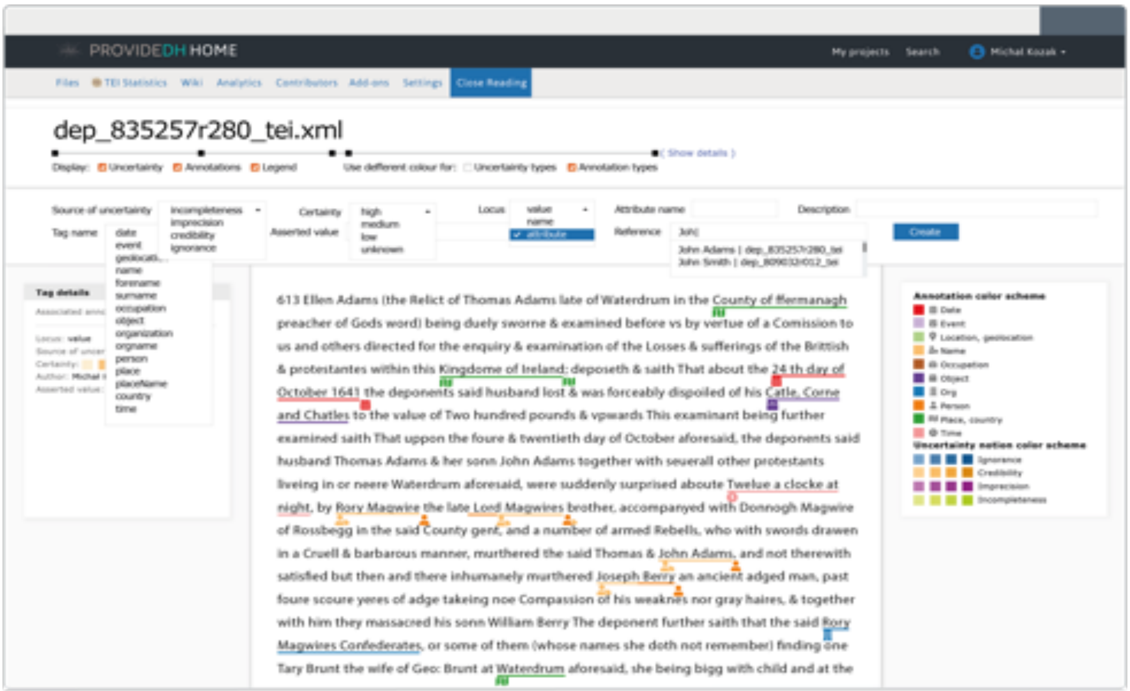
Locating and tracing (certain types of) uncertainty through the evolution of a textual corpus can be done with the use of TEI tags [TEI Consortium 2019]. However, the use of these methods is not a common practice. The motivation of this paper is to address one possible barrier to wider use of these tags by providing a user-friendly interface to collaboratively annotating texts with uncertainty. We propose some minor extensions of the TEI specification that follow from our metrics of uncertainty. The first extension is adding the new "category" attribute to the "certainty" element, required to indicate the source of uncertainty. The second extension is to change the closed list of values of the "locus" attribute to an open list, in order to be able to explicitly indicate the attribute to which uncertainty refers.

Additionally, the authors detected a need to describe the nature and type of uncertainties as well as evaluating the degree of uncertainty the piece of data introduces [Therón et al. 2019].

Our tools on the platform were developed against the background of human-centered design with the focus onto easing uncertainty annotation and visualization, promoting the use of TEI

standards and making uncertainty play a more active role in the research process.

The platform fulfils the common needs of a complete research lifecycle by providing well-known technologies for basic tasks, such as versioning, file management, text editing, and reference resource management.



613 Ellen Adams (the Relict of Thomas Adams late of Waterdrum in the County of ffermanagh preacher of Gods word) being duely sworne & examined before vs by vertue of a Comission to us and others directed for the enquiry & examination of the Losses & sufferings of the Brittish & protestantes within this Kingdome of Ireland: deposeth & saith That about the 24 th day of October 1641 the deponents said husband lost & was forceably dispoiled of his Catle, Corne and Chatles to the value of Two hundred pounds & vpwards This examinant being further examined saith That uppon the foure & twentieth day of October aforesaid, the deponents said husband Thomas Adams & her sonn John Adams together with seuerall other protestants liveing in or neere Waterdrum aforesaid, were suddenly surprised aboute Twelue a clocke at night, by Rory Magwire the late Lord Magwires brother, accompanyed with Donnogh Magwire of Rossbegg in the said County gent, and a number of armed Rebells, who with swords drawen in a Cruell & barbarous manner, murthered the said Thomas & John Adams, and not therewith satisfied but then and there inhumanely murthered Joseph Berry an ancient adged man, past foure scoure yeres of adge takeing noe Compassion of his weaknes nor gray haires, & together with him they massacred his sonn William Berry The deponent further saith that the said Rory Magwires Confederates, or some of them (whose names she doth not remember) finding one Tary Brunt the wife of Geo: Brunt at Waterdrum aforesaid, she being bigg with child and at the

Figure 1. Screenshot of the collaborative TEI editor on the PROVIDEDH Open Science Platform.

The authors aim to get feedback from the TEI community to improve their tools in a generic way and fulfill further needs of the audience.

Bibliography

TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.5.0*. Section 21.1.2 Structured Indications of Uncertainty. http://tei-c.org/release/doc/tei-p5-doc/en/html/CE.html#CECECE

Therón, Roberto at al. 2019. “Conceptualising Uncertainty.” In *Informatics*. Submitted.

The project **Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea** (Schriftkultur des christlichen Äthiopiens: eine multimediale Forschungsumgebung) is a long-term project that aims at creating a digital research environment for the study of the Ethiopian and Eritrean manuscript culture. Digital catalogue entries of manuscripts, text editions, and other relevant records (e.g. certain persons) are encoded in TEI XML. One of the important outcomes of the project is a systematic description of the literary and non-literary texts attested in the Ethiopian and Eritrean manuscript culture. That implies not only encoding, but the establishing of semantically meaningful relations between the texts.

Ethiopian Psalter of the Virgin (Mazmura Dəngəl) is an illustrative example of a text with a complex structure and relations with other texts, which can be precisely described through the encoding in TEI XML. The Psalter of the Virgin is a hymnological composition dedicated to St. Mary. It consists of the Psalms of David, the canticles of the Prophets, and the five sections of the Song of Songs (those texts constituting traditionally the Ethiopian Psalter), as well as additional texts of a much later origin, ascribed to a certain monk named Mazmura Dəngəl from the 15th century and dedicated to St. Mary (to which I refer to as Marian texts). Those Marian texts consist of an opening prayer and 150 short psalms which correspond to and are inspired by the Psalms of David, 15 prayers corresponding to the canticles of the Prophets and five prayers inspired by the Song of Songs. They are usually written directly after the corresponding texts of the Psalter.[1]

The presence of those Marian texts defines the whole work as Mazmura Dəngəl (Psalter of the Virgin). However, it appears that one can't define Mazmura Dəngəl as a collection of the Marian texts. Those Marian texts are always combined with the corresponding texts of the Psalms of David, the canticles of the Prophets, and the Song of Songs.[2] By now there is no evidence for the circulation of the Marian texts of the Psalter of the Virgin on their own. There is evidence for an independent circulation for the opening prayer only, however just in the Psalter manuscripts. Thus, it appears that the Psalter itself constitutes an essential, unavoidable part of the Psalter of the Virgin. This fact might be perfectly reflected by encoding the structure of the Psalter of the Virgin and establishing relations with other texts.

In the project Beta maṣāḥəft any text with an independent circulation is considered a work and thus it gets an individual record with a first-level ID.[3] The Psalms of David, the canticle of the Prophets, and the Song of Songs are attested as text with their independent circulation and ascribed certain IDs in the project. The Marian texts of the Psalter of the Virgin are not attested in their independent circulation, at least by the moment, and thus they can't be provided with their own individual first-level ID. Only the Psalter of the Virgin in which the Marian texts are combined with the texts of the Psalter can be regarded as a work with an independent circulation. The Psalter of the Virgin was ascribed an ID LIT3985Mazmura in the project Beta maṣāḥəft.

By now, I propose the following pattern for encoding of the text structure, which reflects the role of the Psalms of David, the canticles of the Prophets, and the Song of Songs as essential parts of the Psalter of the Virgin and assigns the second level IDs to all Marian texts:

[1] (Sokolinskaia and Pietruschka 2007)

[2] See, for example, the following manuscripts: London, British Library, BL Oriental 621; Portland, Ethiopic Manuscript Imaging Project, Abilene Christian University Codex 2, fols. 2r-158r; Portland, Ethiopic Manuscript Imaging Project, Alwan Codex 2, fols. 1r-94r.

[3] (Liuzzo and Reule 2018)

```
<div type="edition" xml:lang="gez“>

…

<div type="textpart" subtype="Psalm" xml:id="Ps1" n="1">

<label>Psalm 1</label>

<div type="textpart" subtype="PsalmofDavid" corresp="https://betamasaheft.eu/LIT2000Mazmur#Ps1" xml:id="PsD1"/>

<div type="textpart" subtype="PsalmoftheVirgin" xml:id="PsV1"/>

</div>

….

 </div>
```

Typically a psalm of the Psalter of the Virgin consists of a psalm of David and a Marian psalm. And this is reflected in the structure of the Psalter of the Virgin presented here. Each psalm of the Psalter of the Virgin has its own second-level ID, as for example, Ps1 (https://betamasaheft.eu/ LIT3985Mazmura #Ps1). This psalm consists of a Psalm of David, which gets its own second-level ID as part of the Psalter of the Virgin (PsD1) and is linked to the same psalm but as part of an independent text, namely the Psalms of David (https://betamasaheft.eu/LIT2000Mazmur#Ps1). The Marian psalm thus receives its own second level ID.

Using those IDs for encoding of specific manuscripts one can distinguish between the Psalms of David being an independent work and the Psalms of David being part of the Psalter of the Virgin. The same is for the canticles of the Prophets and the Song of Songs. As all three works may constitute a part of the Psalter of the Virgin there is a relation (ecrm:CLP46i_may_form_part_of)[4] established between the Psalms of David, the canticles of the Prophets and the Song of Songs.

To conclude I want to stress the advantages of encoding in TEI XML such complex data as a work or manuscript record, for it allows for very precise defining of the text structure and for avoiding unnecessary generalizations and simplifications, and thus it contributes to better understanding of a complex written tradition.

Bibliographical references:

Liuzzo, Pietro, and Dorothea Reule. 2018. “Beta Maṣāḥǝft Guidelines.” 2018. https://doi.org/10.25592/BetaMasaheft.Guidelines.

Sokolinskaia, Evgenia, and Ute Pietruschka. 2007. “Mäzmurä Dǝngǝl.” In *Encyclopaedia Aethiopica*, edited by Siegbert Uhlig, III:896b–897a. Wiesbaden: Harrassowitz Verlag.

[4] (Liuzzo and Reule 2018)

Reflecting the Influence of Technology on Models of Text in Scholarly Digital Editing

By Julia Josfeld and Grant Leyton Simpson (Georg-August-Universität Göttingen)

Abstract:

There are many resources that aim to help scholars first starting out in the field of digital editing. Courses on XML, XPath, and XSLT, introductions to TEI, tutorials on how to use common software suites, and many more convey a grounded understanding of the technologies involved in this increasingly relevant field. However, one aspect that has received less attention so far is how our application of these technologies influences our approach to and understanding of "text".

In the complex digital medium, the "text" of an edition consists not only of a reproduction of the original source or sources and a number of scholarly apparatuses, but also includes the various layers of infrastructure that make it accessible to potential users. Depending on which choices editors make regarding these infrastructures, the resulting edition will inherit their structural possibilities and limitations, which in turn will dictate constraints on the model of text that can be used. While these underlying influences can sometimes go unexamined, since they appear simply as immutable requirements of the chosen technology, they nevertheless shape the underlying theoretical framework that an editor will work from. It is crucial, therefore, that we use careful discussions of the technologies (e.g. TEI) and their intersection with the material of our editions to elucidate the underlying model possibilities and allow for fruitful and positive decision making.

To illustrate this point, we will show how the application of TEI shaped the model of text we were working from in our digital editing project, the Electronic Corpus of Anonymous Old English Homilies (ECHOE), and how we were able to come to a more nuanced theoretical framework once we started interrogating our base assumptions inherited from this technology.

Keywords:

digital scholarly editing, textual ontology, theory of digital editing

Author Biographies:

Julia Josfeld is a graduate student in English and Medieval Studies at the University of Göttingen, where she also works as a research assistant for the Electronic Corpus of Anonymous Homilies in Old English (ECHOE) project. She received a MSt in Medieval Studies from the University of Oxford, writing on the potential of multilingual digital editions of Medieval Romances. Her main research focus is the theory of digital scholarly editing.

Grant Leyton Simpson is the information technology specialist in digital humanities for the Electronic Corpus of Anonymous Homilies in Old English (ECHOE) project at the University of Göttingen. He received his Ph.D. in English and Information Science from Indiana University. His dissertation studied Old and Middle English DH projects from the 1960s to the present and the objects they produce. His work has appeared in jTEI and will soon appear in Textual Cultures.

Referencing annotations as a core concept of the hallerNet edition and research platform

Peter Daengeli and Christian Forney

ABSTRACT

INDEX

Keywords: publication platform, annotations, models of text, data analysis,

1 May 2019 saw the launch of hallerNet, a platform revolving around prominent actors of the Enlightenment and nature research in eightteenth century Switzerland. HallerNet aims to illuminate the transformation of the early modern *République des Lettres* into the modern scientific community and its discipliniary differentiation by combining digital source editions with a very rich body of prosopographical and bibliographical research data.

2 Whereas the online platform is brand new in its current shape, the underlying metadata was compiled over a span of almost three decades. From the outset, the main focus of the data collection was on Haller's correspondence, the actors related to it and bibliographic information – pertaining to Haller's works, his library but also a vast amount of secondary literature of note –, and resulting in encompassing print publications such as *Repertorium zu Albrecht von Hallers Korrespondenz 1724–*

1777 (Boschung et al. 2002) and *Bibliographia Halleriana* (Steinke and Profos 2004). These endeavors led to a voluminous research database with considerable depth (Steinke 2003), built up between 1991 and 2016 and subsequently transformed into TEI (cf. Recker-Hamm and Stuber 2015, Stuber, Daengeli and Forney 2019). With the onset of a large project on Albrecht von Haller's reviews and letters, funded by the Swiss National Science Foundation (2018–2023), all ca. nine thousand extant reviews by Haller will be edited in conjunction with some eight thousand thematically related letters (on their relationship cf. Stuber 2004).

3 This undertaking, again, relies heavily on earlier research and more specifically on a series of printed editions of Haller's correspondence, which provides the basis for the encoding of more than half of the selected letters. The proposed contribution will discuss the process of the *digitisation* and re-working of such print predecessors. Specifically, the fate of the footnote shall be pondered and the chosen solution in the context of hallerNet editions presented, both on the level of the TEI encoding and the presentational rendering. When developing the data model of these re-editions it quickly showed that porting existing annotations from footnotes in print to footnotes in the digital edition would not leverage the full potential of the new environment. Instead, as much information as possible is attached to references to database objects (persons, institutions, publications, plants and so on). Only critical (philological) annotations and historical information that cannot be related to a database object is retained in footnotes.

4 The implemented model for annotated references is straightforward and basically consists of notes in referencing strings. In the course of the re-edition, information on, e.g., the social position or the place of activity of an actor as it may be given in a legacy footnote is brought over to the respective database object, from where it may be queried also from other occurrences. Consequently, this interweaving of textual data with extensive metadata makes it possible to evaluate this kind of information not only for a single letter, but also for a correspondence as a whole and in doing so to derive and compare social profiles of specific correspondences (Sonntag, Stuber and Forney 2019).

5 A guiding principle of the migration from the (relatively) private database to TEI was to allow for more openness. For one, both the transcribed documents but also the gist of the database objects will be made available in public in a FAIR repository. In addition to this, the data will also be retrievable directly from the platform in a programmatical manner. To this end the database objects are related to authority files wherever possible so that the information may be shared

with other projects and resources. Besides using existing interfaces such as correspSearch and integrating the data with, e.g., HistHub and Metagrid, it will be very interesting to provide access to specific bits of the valuable knowledge contained within the hallerNet platform through nascent interfaces such as prosopogrAPhI.

BIBLIOGRAPHY

Boschung, Urs, Barbara Braun-Bucher, Stefan Hächler, Anne Kathrin Ott, Hubert Steinke,and Martin Stuber. 2002. *Repertorium zu Albrecht von Hallers Korrespondenz 1724–1777* (Studia Halleriana VII/1). Basel: Schwabe.

Recker-Hamm, Ute and Martin Stuber. 2015. *Haller Online – Konzept für den Umbau, Ausbau und die langfristige Sicherung der Haller-/ OeG-Datenbank*: https://files.hallernet.org/public-www/Konzept_Haller_Online_2015_06_08.pdf

Sonntag, Otto, Martin Stuber, and Christian Forney. 2019. "Göttingen and Its Learned Institutions in Albrecht von Haller's European Network: The Example of His Correspondence with Gerlach Adolph von Münchhausen." In *Wissenschaft in Korrespondenzen*, edited by Karsten Engel. Vandenhoeck & Ruprecht 2019 (in print).

Steinke, Hubert. 2003. "Archive databases as advanced research tools: the Haller Project." In *L'edizione del testo scientifico d'età moderna, a cura di Maria Teresa Monti*, edited by Antonio Vallisneri, 191–204. Firenze.

Steinke, Hubert, and Claudia Profos. 2004. *Bibliographia Halleriana. Verzeichnis der Schriften von und über Albrecht von Haller* (Studia Halleriana VIII). Basel: Schwabe.

Stuber, Martin. 2004. "Journal and letter. The interaction between two communication media in the correspondence of Albrecht von Haller." In *Enlightenment, Revolution and the periodical press (Studies on Voltaire and the Eighteenth Century)*, edited by Hans-Jürgen Luesebrink and Jeremy Popkin, 114–41. Liverpool University Press.

Stuber, Martin, Peter Daengeli, and Christian Forney. 2019. "Vom Stellenkommentar zum Netzwerk und zurück: grosse Quellenkorpora und tief erschlossene Strukturdaten auf hallerNet." DHd 2019, 29.05.2019, Frankfurt: https://files.hallernet.org/public-www/presentations/dhd2019/.

AUTHORS

PETER DAENGELI

Peter Daengeli is a researcher at the Cologne Center for eHumanities (University of Cologne) and at the Institute of History, University of Bern (*hallerNet*).

CHRISTIAN FORNEY

Christian Forney is a researcher at the Institute of History, University of Bern (*hallerNet*).

Text Graph Ontology

A Semantic Web approach to represent genetic scholarly editions

Peter Hinkelmanns, University of Salzburg

A model of text as a variant graph can support representing genetic text editions. The proposed model makes it possible to describe the relations between tokens and their relative dependencies in text genesis. The main focus is on the representation of intradocumentary text revisions. Moreover the Text-Graph-Ontology enables the referencing of genetic text editions via the *Semantic Web*. In addition to this ontology, a converter from and to TEI-XML and a web-based viewer and editor will be presented.

Graph, Ontology, Semantic Web, Genetic Text Editions

Text genetic editions are enjoying sustained popularity in the fields of scholarly editions and literary studies. Representatives of recent research projects are the Faust edition (Bohnenkamp, Henke, and Jannidis 2016) or the edition of the works of Arthur Schnitzler (Burch et al. 2016), both of which aim at a complete reproduction of the text genesis. These editions require the reconstruction of complex text genetic processes. Which sequence of tokens forms a specific text state? How can differences between versions be described? The extension of the model of the Text Encoding Initiative by elements required for genetic editions was the subject of a working group that presented its results in a draft (Burnard et al. 2010). Parts of this draft have been incorporated into the TEI Guidelines (TEI Consortium 2013). With TEI P5, complex genetic editions can be realized. However, the underlying structure of the hierarchical graph makes it difficult to reconstruct and compare text gradients, i.e. the evolutionary stages of a text with inline markup. It can of course be done using stand-off and out-of-line annotations, as James Cummings has pointed out (Cummings 2018, 13). This concept for a Text Graph Ontology does not aim to be the next paper criticizing the xml foundation of TEI P5 nor is it trying to compete with the interoperability of TEI encoded texts. The ontology is designed for the specific purpose of implementing a text variant graph using semantic web technologies for the encoding of genetic text editions.

Data models for textual representation

Several project used graphs to represent textual variation in the recent years. They can be divided into methods based on markup and methods focusing on variant graphs. An approach to deal with the problem of overlapping annotations in XML is the proposed markup language ‘General Ordered-Descendant Directed Acyclic Graph’ (GODDAG: Sperberg-McQueen and Huitfeldt 2004, esp. 158). A similar model is the ‘Graph Annotation Format’ (GrAF: Ide and Suderman 2007) which extends the Linguistic Annotation Framework (LAF: Ide and Romary 2006). In GRaF an underlying text is segmented via stand-off nodes, which use the position of characters in the text as reference points. Edges link the annotations with text segments.

The problem of intradocumentary and intertextual variation is addressed by the ‘data structure for representing multi-version texts’ (Schmidt and Colomb 2009). Thy criticize the markup approach of models like GODDAG and GrAF (Schmidt and Colomb 2009, 499) and propose a variant graph where the edges contain the segments of texts shared by multiple variants:

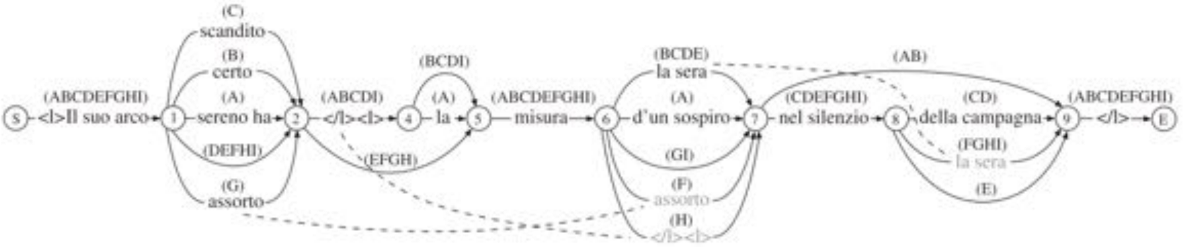


Fig. 1: A variant graph (Figure by Schmidt and Colomb 2009, 502)

The sigils indicate the different variants of the text between different text carriers. Individual tokens of a specific text carrier are not represented.

Collating texts is the focus of the stemmaweb project ("The Stemmaweb Project" 2012–; Andrews and Mace 2013). Similar to Schmidt and Colomb 2009 a variation graph is used to represent versions of a text. The texts are segmented into tokens and form the nodes of the graph. Directed edges show the similarities and dissimilarities between individual text carriers. Undirected edges represent variant relationships like orthographic variation, grammatical variation, lexical variation etc. (Andrews and Mace 2013, 508).

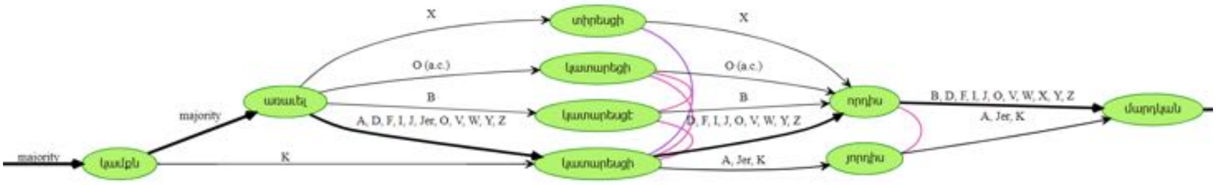

majority
majority
K
X
O (a.c.)
B
A, D, F, I, J, Jer, O, V, W, Y, Z
X
O (a.c.)
B
D, F, I, J, O, V, W, Y, Z
A, Jer, K
B, D, F, I, J, O, V, W, X, Y, Z
A, Jer, K

Fig. 2: Screenshot of the 'Text relationship mapper' of Stemmaweb, showing an extract of Segment 1 of the Chronicle of Matthew (Figure by "The Stemmaweb Project" 2012–)

Efer 2016 has comprehensively described the use of graph databases for the text-oriented Digital Humanities. He proposed 'Kadmos', a layered graph model for textual representation. His model includes the separation into types and tokens:

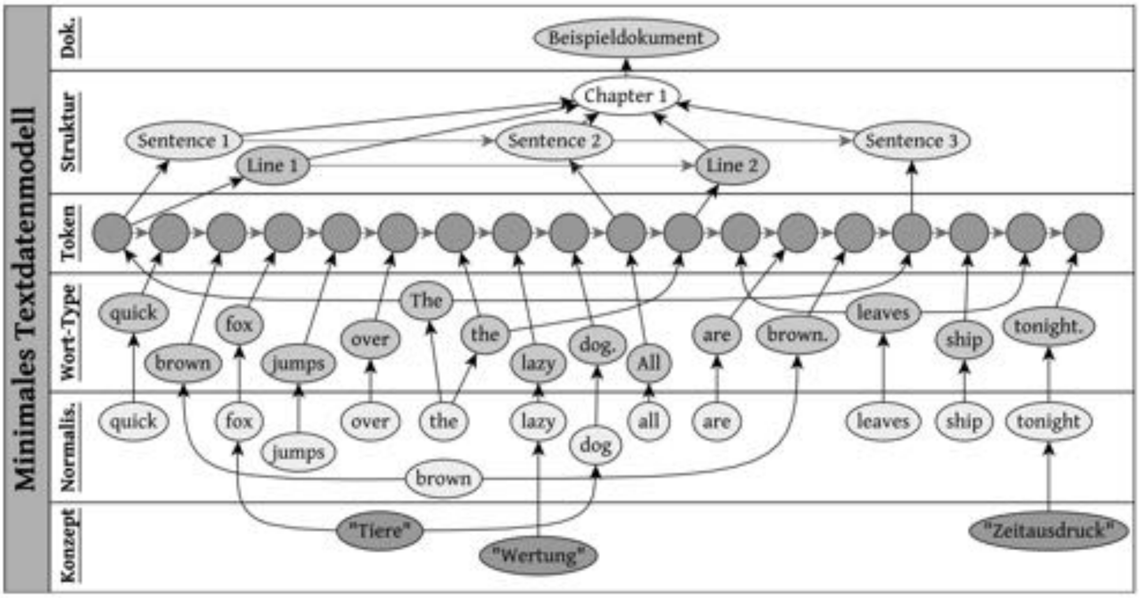


Fig. 3: Schematic representation of instance data sets and links of a short example document with minimalistic text data model (Figure by Efer 2016, 76)

The 'Text as Graph' model (TAG: Haentjens Dekker and Birnbaum 2017) stores the text in nodes of various length (Haentjens Dekker and Birnbaum 2017, §3). A token may be split into several nodes and marked up as a word:

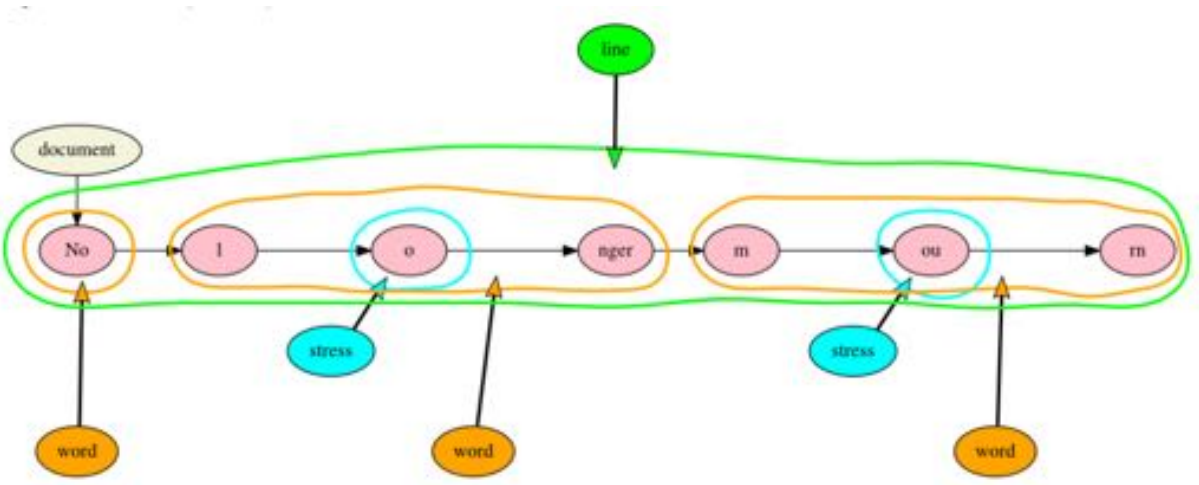


Fig. 4: A simplified poem with word tokenization (Figure by Haentjens Dekker and Birnbaum 2017, Fig. 10)

TAG is at the current stage defined as a data model and not as a syntactic representation (Haentjens Dekker and Birnbaum 2017, §2.1).

This very brief overview of selected data models for textual representation has shown, that multiple models for the representation of texts as graphs exist. The proposed model of this paper is far from state of a stable model and aims to connect the idea of a variation graph with semantic web technologies.

Text graph ontology

The Semantic Web makes information accessible in a machine-readable way using standardized vocabularies and ontologies. A distinction is made between 'nodes', which can be objects or atomic values, and 'edges', which describe the relationship between nodes. A statement always consists of a triple '[subject] - [predicate] - [object]'. Semantic Web technologies enable easy annotating and linking the scholarly genetic edition other resources. The Text Graph Ontology uses the Web Ontology Language (OWL: W3C OWL Working Group 2012) to specify classes and properties.

The first problem which needs to be addressed is the segmentation of a text. There is no agreement between the briefly presented models on what the atomic unit of a text graph is. For the Text Graph Ontology tokens separated by white space should be assumed, with the possibility to extend the model on a sub token level. A diplomatic transcription and various normalization stages can be attached to a token as a string or generatet from a separated character graph.

The proposals for the annotation of text revisions of the Grazer Editionsphilogie are tailored to the needs of mediaeval editions. Hofmeister-Winter 2016 presents a categorization of text revision phenomena. She distinguishes between self-revisions, i.e. interventions in one's own text, and external revisions, which describe the interventions of another hand (Hofmeister-Winter 2016, 10). Self-revisions can be a direct component of text production (immediate revision) or take place at a later point in time (late revision). In their opinion, third-party revisions, on the other hand, take place exclusively in a later revision step as a late revision. Furthermore, the following typology is established by the Graz project (Hofmeister, Böhm, and Klug 2016, 22):

1. **eradication** by bleaching, deletion, blackening, expansion

2. **transformation** resp. transformation by overwriting, addition, reduction
3. **insertion** in all described positions (interlinear, linear, marginal) after eradication by deletion or bleaching, with instruction signs in different shapes, single or paired, as well as gap filling after precautionary recess

To represent these revisions, an edge weighted directed acyclic graph is used. The limitations of RDF make is necessary to construct weighted edges as individual nodes. The base model therefore consists of three node classes: Tokens, Connectors and Borders (the empty start and end nodes of a graph):

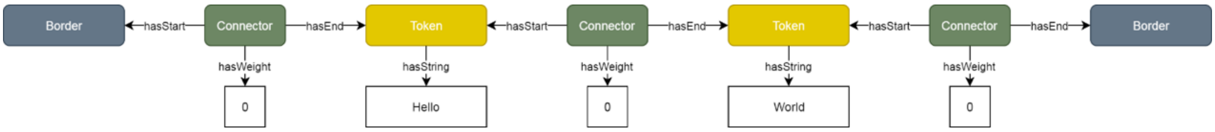


Fig. 5: ‚Hello World‘ as a weighted text graph in RDF

The weight represents the relative order of edges from one token to the next. A substitution would therefore be represented as follows:

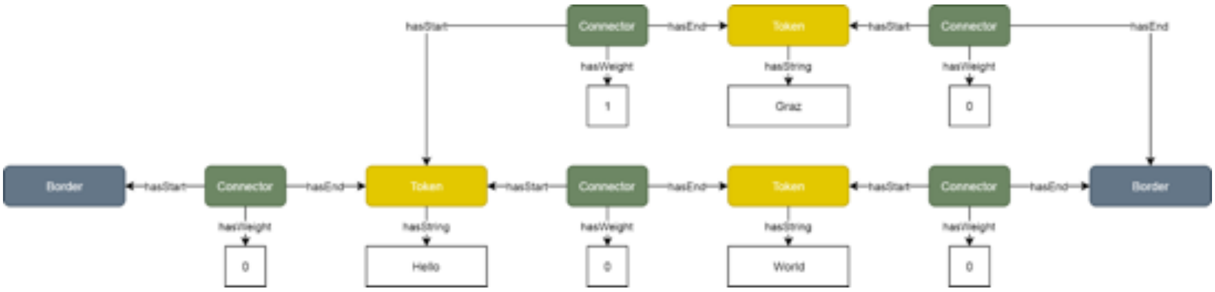


Fig. 6: ‚Hello ~~World~~ Graz‘ as a weighted text graph in RDF

The path following the lowest weight is the first, the past following the highest weight is the last version of the text. The reconstruction of a particular text state can be described as a path through the text. Deletions and additions of text can be seen in the graph accordingly:

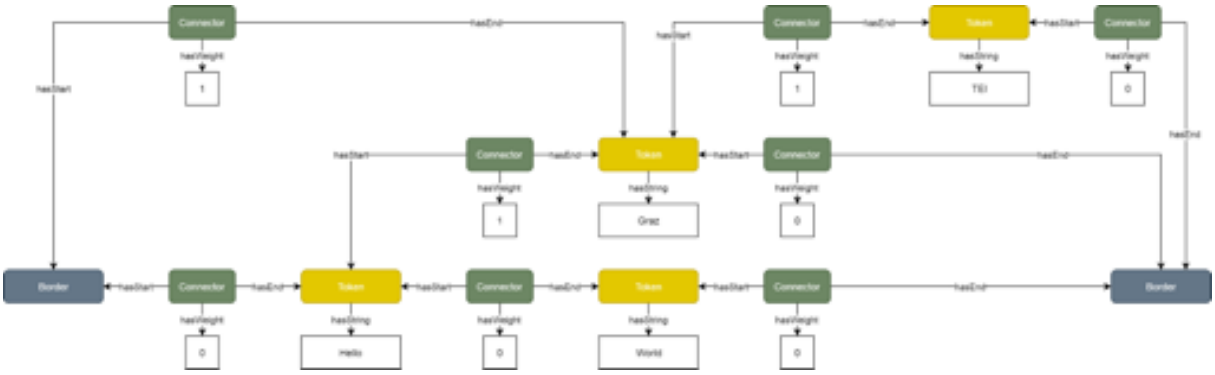


Fig. 7: Transformation, addition and deletion in a weighted text graph

A conversion of this small graph to TEI P5 is possible:

```
<del>Hello</del> <subst><del>World</del><add>Graz</add></subst> <add>TEI</add>
```

To mark specific stages of a text, the Connectors are being referenced to a text stage:

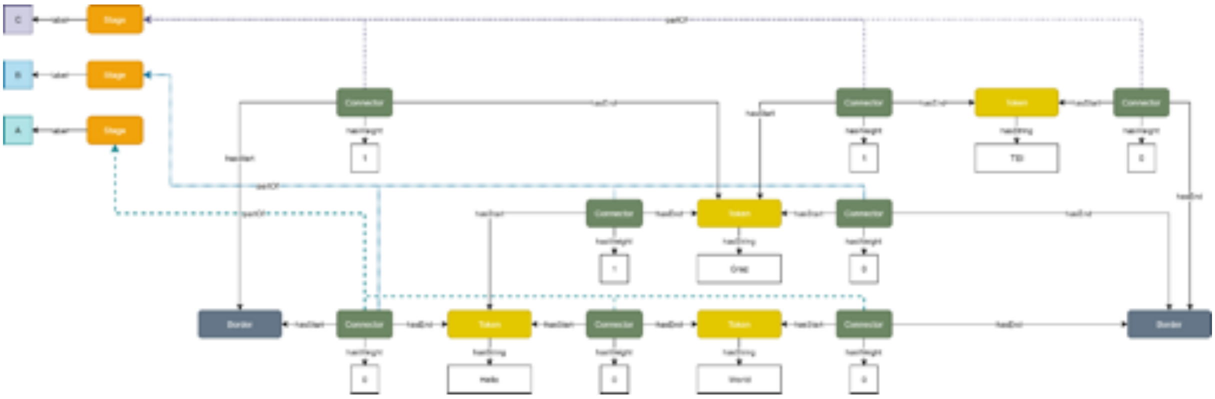


Fig. 8: Text stages realised in a variant text graph

The same variant graph method can be used to describe a token further on character level:

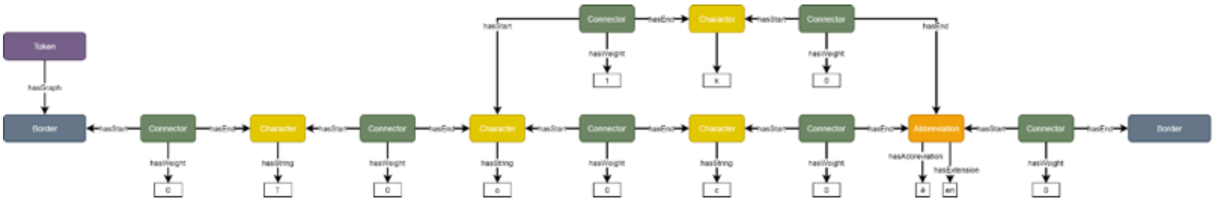


Fig. 9: Graph on character level

This short article shows that semantic web technologies are suitable for representing text variant graphs. The main benefit from using RDF is the interconnectivity with other semantic web ressources. I. e. different text carries of one text could be transcribed by different projects and easily be linked with each other. Annotations on a text can directly point to authority files and vice versa. Tools based on the model are a converter to and from TEI P5 and a viewer/editor based on FLASK. The ontology and the tools will be published on Github. A challenge of this graph approach that has not yet been solved is rule-based validation.

Peter Hinkelmanns, Senior Scientist, University of Salzburg: Middle High German Conceptual Database, peter.hinkelmanns@sbg.ac.at

Peter Hinkelmanns is a senior scientist at the Middle High German Conceptual Database of the University of Salzburg. His research interests include Middle High German lexicography, graph technologies in the digital humanities and historic linguistics.

References

Andrews, T. L., and C. Mace. 2013. “Beyond the Tree of Texts: Building an Empirical Model of Scribal Variation Through Graph Analysis of Texts and Stemmata.” *International Journal of Human-Computer Studies* 28 (4): 504–21. doi:10.1093/llc/fqt032.

Bohnenkamp, Anne, Silke Henke, and Fotis Jannidis. 2016. “Johann Wolfgang Goethe: Faust: Historisch-Kritische Edition.” 2. Beta-Version. Accessed April 26, 2017. http://beta.faustedition.net/.

Burch, Thomas, Stefan Büdenbender, Kristina Fink, Vivien Friedrich, Patrick Heck, Wolfgang Lukas, Kathrin Nühlen et al. 2016. “Text[ge]schichten: Herausforderungen Textgenetischen Edierens Bei Arthur Schnitzler.” In *Textgenese Und Digitales Edieren: Wolfgang Koeppens „Jugend“ Im Kontext Der Editionsphilologie*, edited by Katharina Krüger, 87–105. Editio / Beihefte, 40. Berlin, Boston: de Gruyter.

Burnard, Lou, Fotis Jannidis, Elena Pierazzo, and Malte Rehbein. 2010. “An Encoding Model for Genetic Editions.” Accessed March 25, 2019. http://www.tei-c.org/Activities/Council/Working/tcw19.html.

Cummings, James. 2018. “A World of Difference: Myths and Misconceptions About the TEI.” *Digital Scholarship Humanities* 6:i63. doi:10.1093/llc/fqy071.

Efer, Thomas. 2016. “Graphdatenbanken Für Die Textorientierten E-Humanities.” Dissertation, Universität Leipzig.

Haentjens Dekker, Ronald, and David J. Birnbaum. 2017. “It’s More Than Just Overlap: Text as Graph.” In *Proceedings of Balisage: The Markup Conference 2017*, edited by B. T. Usdin, Deborah A. Lapeyre, James D. Mason, C. M. Sperberg-McQueen, and Norman Walsh. Balisage Series on Markup Technologies: Mulberry Technologies, Inc.Rockville, Maryland.

Hofmeister, Wernfried, Astrid Böhm, and Helmut W. Klug. 2016. “Die Deutschsprachigen Marginaltexte Der Grazer Handschrift UB, Ms. 781 Als Interdisziplinärer Prüfstein Explorativer Revisionsforschung Und Editionstechnik.” *Editio* 30 (1): 14–33. doi:10.1515/editio-2016-0002.

Hofmeister-Winter, Andrea. 2016. “Beredte Verbesserungen.” *Editio* 30 (1): 1–13. doi:10.1515/editio-2016-0001.

Ide, Nancy, and Laurent Romary. 2006. “Representing Linguistic Corpora and Their Annotations.” In *Proceedings of the Fifth Language Resources and Evaluation Conference: LREC 2006*, edited by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias: Genua.

Ide, Nancy, and Keith Suderman. 2007. “GrAF: A Graph-Based Format for Linguistic Annotations.” In *Proceedings of the Linguistic Annotation Workshop: Held in Conjunction with ACL 2007*, edited by Association for Computational Linguistics, 1–8. https://www.aclweb.org/anthology/W07-1501. Accessed July 30, 2019.

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. doi:10.1016/j.ijhcs.2009.02.001.

Sperberg-McQueen, C. M., and Claus Huitfeldt. 2004. “GODDAG: A Data Structure for Overlapping Hierarchies.” In *Digital Documents: Systems and Principles*, edited by Peter King and Ethan V. Munson, 139–60. Berlin, Heidelberg: Springer.

TEI Consortium. 2013. “TEI P5: Guidelines for Electronic Text Encoding and Interchange.” Version 3.6.0. Accessed July 30, 2019. https://www.tei-c.org/Vault/P5/3.6.0/doc/tei-p5-doc/en/html/.

“The Stemmaweb Project: Tools and Techniques for Empirical Stemmatology.” 2012–. Accessed July 30, 2019. https://stemmaweb.net/.

W3C OWL Working Group. 2012. “OWL 2 Web Ontology Language: Document Overview (Second Edition).” W3C Recommendation 11 December 2012. Accessed June 08, 2018. https://www.w3.org/TR/owl2-overview/.

Long Paper Proposal: TEI 2019
Jennifer Roberts-Smith and Joey Takeda
with Janelle Jenstad, Mark Kaethler, and Toby Malone

Reconceiving TEI models of theatrical performance text with reference to promptbooks

This paper explores and suggests a revision to the conceptual model of “dramatic text” currently evidenced in TEI discourse. Our critique arises out of the challenges our research team has encountered developing an encoding protocol for promptbooks in the collection of the Canadian Stratford Shakespeare Festival Archives. We have argued elsewhere that promptbooks—by which we mean the promptbooks defined by Charles H. Shattuck (1965) as the “book[s] actually used by prompters or stage managers in conducting performances” (1965, 5)—are ontologically distinct from other forms of theatrical performance texts in the sense that “a promptbook records a series of utterances to which intended performance events are mapped in a relative temporal sequence” (Roberts-Smith et al. 2019).

In this paper, we work through the implications of that claim for the TEI Guidelines, and suggest a reconceptualization of TEI’s discursive model of theatrical performance texts based on:

1. long-acknowledged limitations of the existing Guidelines to address the complex ontologies of a “dramatic work” (Mylonas and Lavagnino 1995);
2. a summary of more recent theoretical work on the ontologies of dramatic and theatrical/performance texts (including frameworks proposed by Osborne 1996; Clopper 2001; Erne 2003; Schafer 2006; Kidnie 2009; Worthen 2005, 2010, 2011; Werstine 2012; Roberts-Smith et al. 2013; Griffin 2018); and
3. a survey of existing TEI Guidelines for Performance Texts (TEI 7), TEI listServ discussions (with a focus on the 2001 debate among Tobin Nellhaus, Syd Bauman, Sperberg-McQueen, and others[1]), and other TEI extensions, including the Music Encoding Initiative (MEI).

Our central argument is that the tension between event and text that Mylonas and Lavaigno see as problematic in encoding “dramatic works” (1995) applies neither to performance texts traditionally understood as such nor to promptbooks. A more functional conceptualization in TEI would accommodate dramatic, theatrical/performance, and prompt texts as three distinct ontologies. We offer a short list of sample modifications to the TEI Guidelines for discussion.

References

Bauman, Syd. (2001 April 12). “Re: TEI and drama”. TEI Electronic ListServ. https://listserv.brown.edu/archives/cgi-bin/wa?A2=TEI-L;2973394d.0104

Clopper, Lawrence (2001). Drama, Play, and Game: English Festive Culture in the Medieval and Early Modern Period. Chicago: University of Chicago Press.

[1] Note that full thread is not available via the TEI listServ as it appears some of the discussion happened “off-list”. See Nellhaus’ initial post https://listserv.brown.edu/cgi-bin/wa?A2=TEI-L;efa95b0.0103 and Bauman’s reply: https://listserv.brown.edu/cgi-bin/wa?A2=TEI-L;2973394d.0104

Erne, Lukas (2003). *Shakespeare as Literary Dramatist*. Cambridge: Cambridge University Press.

Griffin, Andrew (2018). “Text, performance, and multidisciplinarity: On a digital edition of *King Leir*.” In *Shakespeare’s Language in Digital Media: Old Words New Tools*, ed. Janelle Jenstad, Mark Kaethler, and Jennifer Roberts-Smith. 84-104. Abingdon: Routledge.

Kidnie, M. J. (2009). *Shakespeare and the Problem of Adaptation*. Abingdon: Routledge.

Lavagnino, John, and Elli Mylonas. (1995). “The Show Must Go on: Problems of Tagging Performance Texts.” *Computers and the Humanities* 29: 113-121.

Music Encoding Initiative: Guidelines. 4.0.1. *April 12, 2019. Music Encoding Initiative. https://music-encoding.org/guidelines/v4/content/*.

Osborne, Laurie E. (1996). “Rethinking the Performance Editions: Theatrical and Textual Productions of Shakespeare.” In *Shakespeare, Theory and Performance*, ed. James Bulman, 168-86. Abingdon: Routledge.

Roberts-Smith, Jennifer, Mark Kaethler, Toby Malone, Liza Giffen, Janelle Jenstad, Martin Holmes, and Joseph Takeda (2019). “Tagging Time and Space: TEI and the Canadian Stratford Festival Promptbooks.” *Digital Studies/Le champ numérique* 9.1: https://doi.org/10.1017/CBO9781139103978

Roberts-Smith, Jennifer, et al. (2014). “Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET).” *Digital Humanities Quarterly* 7.3.

Schafer, Elizabeth (2006). “Performance Editions, Editing, and Editors.” *Shakespeare Survey* 59: 198-212.

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. 3.5.0. January 29, 2019. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Werstine, Paul. 2012. *Early Modern Playhouse Manuscripts and the Editing of Shakespeare*. Cambridge: Cambridge University Press.

Worthen, W. B. (2005). *Print, and the Poetics of Modern Drama*. Cambridge: Cambridge University Press.

Worthen, W.B. (2010). *Drama: Between Poetry and Performance*. Chichester: Wiley-Blackwell.

Worthen, W.B. (2011). Intoxicating Rhythms: Or, Shakespeare, Literary Drama, and Performance (Studies). *Shakespeare Quarterly* 62.3 (fall): 303-339.

Using Github and its Integrations to Create, Test, and Deploy a Digital Edition

Authors: Joseph Takeda, Sydney Lines

This paper stems from the ongoing work by the Winnifred Eaton Archive (WEA), which seeks to compile, transcribe, and encode the extant archive of Chinese-Canadian author Winnifred Eaton (1875–1954). While there are many frameworks for rendering TEI online (including the TEI Stylesheets, TEI Boilerplate, and CETEIcean), the WEA, like many other projects housed at institutions without a dedicated digital humanities infrastructure, struggled to find a framework for testing, deploying, and publishing the project as a whole; Omeka and Wordpress were offered as solutions, but these frameworks are limited in their capacity to handle TEI-encoded XML. Following the best practices outlined by The Endings Project (Carlin 2018) and inspired by the recent turn to static sites for digital editions (Holmes 2017; Viglianti 2017), 'minimal editions' (Gil 2015; Sayers 2016; Gil 2017), and web publishing at large (Rinaldi 2015), we arrived at the following workflow:

- Store all content on Github
- Integrate Travis-CI with repository to build and validate products
- Use Travis to deploy to a separate Github repository that deploys content using the Github pages environment

This paper thus forwards the above method as a wide-ranging, affordable solution for creating digital projects in TEI (it is entirely free, minus the optional costs of oXygen XML Editor and a domain name) that is sustainable and robust as it leverages existing technologies that are ubiquitous and well-documented. This process is also highly extensible and can be used in concert with existing TEI publishing solutions, like TEI Boilerplate, to create sustainable and archivable static digital projects that are not beholden to the structural limits of pre-existing content management systems. Our paper explains the major benefits of this approach, which include affordability, sustainability, and adaptability, as well as suggests the potentials of this approach across various pedagogical and scholarly publishing workflows.

TEI Lex-0: a good fit for the encoding of the Portuguese Academy Dictionary?

Ana Salgado(1), Rute Costa(1), Toma Tasovac(2)

(1) NOVA CLUNL, Universidade NOVA de Lisboa
(2) Belgrade Center for Digital Humanities, Serbia

anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt; ttasovac@humanistika.org

In this paper, we report on the encoding of the Portuguese Academy Dictionary using TEI Lex-0. We demonstrate how we applied this new baseline format for lexical data to mark up ‘special entries’ in the dictionary: part-of-speech homonyms (*capital*1, *capital*2, *capital*3), etymological homonyms (*cota*1, *cota*2), homographs (*lobo*1 /ó/, *lobo*2 /ô/), spelling variants (*ouro, oiro*), trademarks (*donut*), entries that have a different meaning in the plural (*antepassados*), and lexical variants (*missanga, miçanga*). Even though TEI Lex-0 reduces the number of TEI elements that can be used to describe entry-like objects from five (<entryFree>, <entry>, <superEntry>, <hom> and <re>) to only one (<entry>), our work shows that TEI Lex-0 is fully capable of representing the complexities of the entry structure of the Portuguese Academy Dictionary. Furthermore, we argue that this simplified array of elements can lead to more coherent and more legible encoding without sacrificing its semantic expressivity. In addition to justifying our concrete encoding choices, we will describe the process of converting our data from TEI to TEI Lex-0 and the documentation of the differences between our original TEI encoding and the TEI Lex-0 version. As of this writing, TEI Lex-0[1] is still a work in progress. This paper is therefore intended as both a contribution to and a commentary on the efforts of the TEI Lex-0 group.

Keywords: dictionary encoding, general monolingual dictionary, lexical database, lexicography, TEI Lex-0.

Acknowledgements
Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2019, and by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

References

Dictionary

Dicionário da Língua Portuguesa Contemporânea, 2001, João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo. New digital edition under revision.

[1] https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0

Other literature

Bański, P., Bowers, J., Erjavec, T. "TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms". In *Electronic Lexicography in the 21st Century*: Proceedings of eLex 2017 Conference, pp. 485-94, 2017.

Bowers, J., Romary, L. "Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec". In *Dictionaries: Journal of the Dictionary Society of North America, Dictionary Society of North America*, 39(2), pp. 79-106, 2018.

Budin, G., Majewski, S., Morth, K. "Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries". In *Journal of the Text Encoding Initiative* 3, http://jtei.revues.org/522, 2012.

Romary, L. "TEI and LMF crosswalks". In *Digital Humanities*: Wissenschaft vom Verstehen. Berlin: Humboldt Universität zu Berlin, 2015.

Romary, L., Tasovac, T. "TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources". In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pp. 274-275, https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf, 2018.

Simões, A., Almeida, J. J., Salgado, A. "Building a Dictionary using XML Technology". In *5th Symposium on Languages, Applications and Technologies (SLATE'16)*, vol. 51 of Open Access Series in Informatics (OASIcs), pp. 14:1-14:8. Germany: Dagstuhl. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. DOI: http://dx.doi.org/10.4230/OASIcs.SLATE.2016.14, 2016.

Tasovac, T. "Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities". In *Digital Humanities 2010*, pp. 254-256, 2010.

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* [Version 3.5.0]. [Last updated on 29th January 2019, revision 3c0c64ec4]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ ([28.07.2019]).

Biography

Ana Salgado: Doctoral candidate on Translation and Terminology at NOVA FCSH, researcher at Center of Linguistic of the NOVA University (CLUNL). Lexicographer, coordinator of the Spelling Vocabulary of the Portuguese Language and the new Dictionary of the *Academia das Ciências de Lisboa*. Her research interests include Lexicography, Digital Humanities and data modelling. Her PhD research project focuses on the treatment of terms in general language dictionaries, which aims to combine lexicographic and terminological methodologies in the planning of the macro- and microstructure of monolingual dictionaries in order to improve the quality and interoperability of lexical databases.
http://clunl.fcsh.unl.pt/equipa/ana-salgado/

Rute Costa: Tenured Associate Professor of Linguistics with Habilitation – Lexicology, Lexicography, Terminology at NOVA FCSH. President of the Center of Linguistic of the NOVA University (CLUNL). President of ISO/TC37/SC2 'Terminology workflow and language coding' Former president of the European Association for Terminology (EAFT). At NOVA FCSH, she lectures on Terminology, Terminology Theories, Terminology and Ontologies, among other subjects. She director of the PhD Program in Translation and Terminology (NOVA &UAveiro). She is affiliated with the H2020 funded project European Lexicographic Infrastructure (ELEXIS) and member of the Value4Helath Lab (NOVA).
http://clunl.fcsh.unl.pt/equipa/rute-costa/

Toma Tasovac: Director of the Belgrade Center for Digital Humanities (BCDH) and Director of the Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU). His areas of scholarly interest include lexicography, data modeling, TEI, digital editions and research infrastructures. He is one of the authors of TEI Lex-0, an emerging community standard for encoding lexicographic data and is currently also affiliated with the H2020-funded project European Lexicographic Infrastructure (ELEXIS).
http://humanistika.org

Ermenegilda Rachel Mueller, University of Iceland
ermenegilda.muller@gmail.com
Conftool ID of abstract: 177

A TEI customization for the description of paper and watermarks

1. Summary

TEI offers a large set of tools to describe material features of written documents. However, these tools are not yet sufficient to produce comprehensive, structured descriptions of paper, and do not reflect the standards of paper historians. The present contribution consists of a TEI-P5 customization for the description of paper and watermarks. This customization is designed to let users choose the level of detail they wish to achieve in such descriptions while maintaining standardized terminology, criteria for data recording, and structure.

2. Scope and purpose

Being able to record standardized paper and watermark data in TEI would not only be useful to projects that focus on the study of paper. It would also be a significant asset for projects that use TEI to describe paper documents in general. Indeed, watermarks can be used to date manuscripts and printed books (see for instance Stevenson, 1967). Moreover, watermark motives are indicators of paper quality, and can thus be used to assess the financial investment represented by the production of a given manuscript or printed book (see for instance Busonero et al., 2001). Finally, they bear witness to the geographical origin of paper, and can thus inform where documents were made or the commercial routes through which the paper used to produce them was obtained (see for instance Churchill, 1935). Therefore, they are key to determining the historical, geographical and socio-economical contexts in which documents were produced. Including this information in TEI descriptions of primary sources would facilitate collaborations between specialists - for instance manuscript scholars, book historians and paper scholars - and thus benefit our knowledge of the origin and production context of paper manuscripts and early printed books.

The present TEI customization is modeled on the international standard for the description of paper, watermarks and paper molds in relational databases, IPHN 2.1.1 (IPH, 2013).

The purpose of IPHN is to standardize the recording of data concerning historical and modern papers, with or without watermarks, in order to ensure the compatibility of the different digital databases of papers and watermarks. It offers a tagset and a database structure for the registration of all the features of papers and watermarks that contribute to their identification, and a standardized terminology for watermarks motives. At the present stage, IPHN is the only existing standard for papers and watermarks registration in digital databases, and it is not yet widely applied. Most digital databases of watermarks use different digitizing methods, different criteria for the registration of data, and different terminology. For these reasons, their interoperability is limited, despite the efforts to have been made in this direction (see especially Bernstein Project, 2019). Adapting the IPHN standard to TEI-P5 significantly contributes in fixing this problem in three ways:

- First of all, since TEI is used in a majority of digital catalogues covering paper documents, it allows a larger number of professionals - scholars of manuscripts and early printed texts, archivists, librarians etc. - to record paper and watermark data.
- For the same reason, it makes paper and watermark data more accessible and allows scholars to better localize paper types and watermarks in collections of paper documents.
- Finally, since TEI is designed to be highly stable and interoperable, using it for paper and watermarks data ensures that this data is recorded and encoded in a standardized manner.

3. Structure and content of the customization

The present customization combines new elements based on the parameters listed in the IPHN standard with the official TEI modules **msdescription**, which is suitable for the physical description of all text-bearing objects, **namesdates**, which provides the necessary elements for the description of persons, places and organizations, and can therefore be used to enter information about paper makers and paper mills, and **linking**, in order to link the information together. It consists of two modules:

- One module for the description of papers and watermarks themselves.
- One module for the description of paper molds and the registration of information concerning paper mills and paper makers.

The first custom module allows users to nest descriptions of one or more paper types and/or watermarks in the **support** and/or **watermark** elements of the **msdescription** module. These descriptions contain the following information:

- The locus of a given paper type within the quire structure of the primary source that is described.
- The physical features of this paper type (aspect, dimensions, state of conservation) and, when one can assess it, its composition and mode of production.
- The number of chain lines (i.e. the vertical lines that are imprinted into the paper by the mold) and the distance between them.
- The density of laid lines (i.e. the horizontal lines that are imprinted into the paper by the mold).

- The position of the watermark, in relation to the paper sheet and to the chain lines.
- The watermark motive, using the typology defined in IPHN 2.1.1.

The second custom module allows users to enter descriptive information about a paper molds (i.e. chain lines, laid lines and watermark position and motive), geographical and historical information about the paper mill that used it, and prosopographic information about the paper maker. It combines the customization's new elements for paper and watermarks description and the elements provided in the **namesdates** module for the recording of historical and biographical information.

The descriptions of papers and watermarks within primary sources and the descriptions of paper molds are meant to be registered in different files. Both custom modules include the official **linking** module so that users can link them together. This entails that files containing paper mold descriptions are stored in a separate database. However, since identifying paper molds can only be achieved by recording paper and watermarks data in a very large number of manuscripts or printed books, it cannot be done for small corpora or collections.

Therefore, making paper mold description files is optional, and the customization can simply be used to include paper and watermarks data in the description of primary sources without necessarily linking them together.

4. Cost and time investment

The customization is designed to give projects that wish to use it maximum freedom in terms of the time and money they wish to invest in the recording of paper and watermark data. For this purpose, it has the two following features:

- The two custom modules are flexible enough to allow users to record paper and watermark data at the level of detail that they wish to achieve while preserving standardized terminology and structure. The only new elements that are required are those that represent features of paper types and watermarks that are both absolutely necessary to their identification and observable with minimal means (see below). This allows projects to decide how much time they wish to invest in the recording of paper and watermarks data, and to determine if they need to hire staff for this purpose.
- The customization requires that users indicate the method used to collect paper and watermarks data, so that scholars who consult the description are aware of how much information could be collected with the method used, and how accurate this information can be. Very cheap and minimalistic methods can be used to collect such data with satisfying results. Nowadays, the most common of them is the use of a fibre optic light sheet to observe chain lines, laid lines and watermarks. Such a device can be found in the vast majority of

conservation workshops of manuscripts and rare books collections, which means that using it does not entail additional material cost. The information collected with this method is, in many cases, already sufficient for the identification of watermarks. Much more precise methods currently exist - for instance, radiography - and projects that wish to record paper and watermarks data may want to invest in them in order to perform more detailed and accurate observations. This is, however, entirely up to them, and not required for the use of the present customization.

5. Outcomes

The descriptions of paper types, watermarks, and paper molds created with the present customization can be used for various purposes, depending on how detailed they are and for how many items (paper manuscripts and/or printed books) they have been made. At a minimal level of detail and for a minimal number of described items, they allow users of digital catalogues to localize watermarks within collections and within collection items, know the basic features of the paper that contains them, and compare them with the items recorded in other paper and watermarks database, with the objective of identifying them. The more items are described, even at a minimal level of detail, the more likely it is to identify paper types and watermarks by mining the database, and to be able to produce paper molds descriptions. The more detailed the descriptions are, the more accurate the identification of paper types and watermarks will be. Finally, once paper types and watermarks have been identified, they can be used to date and contextualize the items in which they were found.

How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures

Mohamed Khemakhem[1,2,3], Ioana Galleron[4], Geoffrey Williams[6], Laurent Romary[1], Pedro Ortiz Suárez[1,5]

1. Inria-ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities
2. UPD7 - Université Paris Diderot - Paris 7
3. CMB - Centre Marc Bloch, Berlin
4. Université Sorbonne-Nouvelle
5. Sorbonne Université
6. Université Grenoble Alpes

In the last decade, OCR progress has triggered a massive trend towards the digitisation of legacy documents, with several Digital Humanities projects[1][2][3] exploring means for structuring retro-digitised dictionaries. However there is a lack of awareness of the impact of the OCRs quality on the information extraction process. In this work, we shed light on the relationship between these two steps through experiments carried out with a TEI-based system for automatic parsing of dictionaries.

Our work concerns "the Basnage", a complex dictionary resulting from the complete revision and enlargement in 1701 of the 'Dictionnaire Universel' of Abbé Furetière, initially published in 1690. In order to obtain an XML/TEI version of this work, we use GROBID-Dictionaries [1,2], a machine learning system for cascade parsing and extraction of TEI structure in dictionaries. The tool's models have been tested on different categories of entry based documents with lexical and encyclopedic content.
We used two differently OCRied versions of the first volume of the Basnage[4] following the process described in an earlier experiment [3] which relies on the power of iterative training of HTR models of Transkribus[5] framework:

- **Sample 1:** created using an HTR model trained with 28 pages and a low image quality document

[1] https://basnage.hypotheses.org/
[2] https://elex.is/
[3] http://nenufar.huma-num.fr/presentation/
[4] https://archive.org/details/b30455376_0001/page/n27
[5] https://transkribus.eu/Transkribus/

- **Sample 2:** created using an HTR model trained with 108 pages and high image quality document

The results of our experiment are as follows:

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<ab>	99.95	99.95	99.75	81.48	73.33	77.19
<fw type="footer">	100	76.47	86.67	84.62	91.67	88
<fw type="header">	92.59	80.65	86.21	100	90	94.74

Table 1: Field Level Evaluation of the Dictionary Segmentation Model

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<dictScrap>	81.82	85.71	83.72	100	90	94.74
<entry>	85.85	80.53	83.11	89.47	91.07	90.27
<pc>	92.59	96.15	94.34	93.75	97.56	95.62

Table 2: Field Level Evaluation of the Dictionary Body Segmentation Model

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<etym>	87.5	60	71.19	73.68	71.79	72.73
<form>	94.44	92.73	93.58	92.24	96.4	94.27
<pc>	90.91	69.44	78.74	88.97	80.13	84.32
<re>	33.33	9.09	14.29	55.56	22.73	32.26
<sense>	67.65	59.28	63.19	77	76.65	76.84
<xr>	100	80	88.89	100	100	100

Table 3: Field Level Evaluation of the Lexical Entry Model

Conclusion

Two main conclusions could be drawn from these experiments. First, the OCRisation process has an important impact on the performance of the automatic parsing of TEI structures. The impact becomes more significant for extracting more granular information. Second, the models implemented in GROBID-Dictionaries are resistant to the noise introduced by the OCR system opening perspectives for exploiting more available digitised material.

References

1. Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *electronic lexicography, eLex 2017*, Sep 2017, Leiden, Netherlands.
2. Mohamed Khemakhem, Axel Herold, Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX workshop at LREC 2018*, May 2018, Miyazaki, Japan. 2018.
3. David Lindemann, Mohamed Khemakhem, Laurent Romary. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. *European Association for Digital Humanities (EADH) Conference*, EADH, Dec 2018, Galway, Ireland.

Biographies

Mohamed Khemakhem is a PhD candidate at Inria, team ALMAnaCH (Paris), Paris 7 University and Centre Marc Bloch (Berlin). His research is focused on parsing lexical and encyclopedic legacy resources using standard-based machine learning models.

Ioana Galleron is a professor of French literature and Digital Humanities at Sorbonne-Nouvelle and UMR 8094 LATTICE of CNRS. She works on computer assisted literary analysis.

Geoffrey Williams is a Professor of Applied Linguistics at the University of South Brittany and researcher at UMR 5316, Litt & Arts at the University Grenoble Alpes. He is a e-lexicographer and leads the ANR BasNum project.

Laurent Romary is senior researcher at Inria, team ALMAnaCH and works on data modelling and standards in humanities computing.

Pedro Ortiz Suárez is a PhD candidate at Inria, team ALMAnaCH (Paris) and Sorbonne Université. His research is focused on enriching lexical and encyclopedic legacy resources using deep learning models.

What is a Line? Encoding and Counting Lines in Linked Early Modern Drama Online

What we consider to be a line, how we count and number them, and how we encode them so that they can be analyzed and counted are deeply significant and yet fraught decisions. Lines numbers are a key critical metric in Early Modern Drama. Lines are the dominant measure of play and role length. Major critical arguments depend on whether and how much a character speaks in verse or prose; at the same time, editorial relineation is often based on the very arguments that in turn depend on lineation. Complicating the matter is that characters do not always speak in full lines. Part lines are hard to identify as prose or verse. Multiple short utterances by characters speaking in succession may be set on one typographical line. Complicating the mise-en-page still further is the fact that prose was often set as verse and vice versa to fit text to page space, as well as the fact that, as in modern typesetting and digital interfaces, early modern prose lines break where the page or column width demands. Line numbers are a common output of a scholarly edition. The 1623 first folio of Shakespeare's plays has even been subjected to a copyrighted canonical numbering system (Through Line Numbers). Yet all editorial theory since McGann has emphasized the instability of such textual features. Despite their utitility for citation and analytical purposes, line numbers can differ wildly between editions, precisely because of the difficulty of establishing lines and the typographical variations between prose in different formats. The TEI offers three ways of encoding lines: we can locate its beginning <lb>, describe it as a topographic line <line>, or identify it as a line of verse <l>. We discuss the relative merits and implicit critical claims of each encoding method for describing and counting lines. We discuss the tension between citable and fluid texts, outline how we use the @ed and @edRef attributes in our lineation, and introduce our prototype whereby various lineation systems are interoperable. The LEMDO (Linked Early Modern Drama) prototype allows for canonical line identifiers, but does not foreclose other possibilities; in fact, LEMDO proliferates @xml:ids so that projects within the LEMDO ecosystem can make project-level decisions about what constitutes a line. Finally, we gesture towards future reconciliation of lineation systems through linked-open data.

Authors: Janelle Jenstad, Joseph Takeda, Brett Greatley-Hirsch, James Mardock

Growing collections of TEI texts: Some lessons from SARIT

Patrick McAllister (patrick.mcallister@oeaw.ac.at)

There is no certainty as to the size that a corpus of Indic texts would have, but it is certain that this is due, mainly, to its sheer extent, rather than to other factors like the frequent reappearance of lost works or of important new witnesses to existing ones. Any fancy about the constitution of a corpus of Indic texts, however narrowly defined, is quickly sobered by the recognition that such a project lies far beyond the capacity of a single generation of scholars.

Nevertheless, some attempts towards its realization do exist, as do very different ideas of how such a growing collection should be designed and how it could be maintained. The problems that all such attempts must overcome are not only technical, but often practical. For example, the common expectation that texts in a collection should be consistent at least in their formal characteristics can easily conflict with the highly specialized and sometimes changing interests of the scholars producing an edition of a work, or also with the fact that individual sections of a single work are often edited by different scholars working with varying methods.

This talk will reflect on the choices made in SARIT (http://sarit.indology.info), which attempts to provide an environment for an expanding collection of Indic texts. The proposed solution was, and with some modifications still is, to design a repository that, over the very long term, can increase and deepen in such a way that new contributions either improve the existing material, or expand the collection with new material without disrupting the general integrity and basic standards of such a collection.

Towards larger corpora of Indic texts: For now, minimize metatext

Himal Trikha (himal.trikha@oeaw.ac.at)

The Digital Corpus of Vidyānandin’s works (DCVW) is an ongoing collection of digital text resources for the works of a 10th century Sanskrit author. The resources are assembled and maintained in the context of my Indological research specialization, i.e., the history of an Indian philosophical tradition. A web interface (http://www.dipal.org/dcvw) allows to access the resources and, to some extent, modify them.

The digital resources are XML-files that are processed by a bundle of technologies in order to pursue specific research interests: search for text strings, identification of dialogic or intertextual elements, differences between attestations etc. In this context, the quality of the results depends on the quality of the resource files, which are assessed by three basic criteria: (a) status of the separation of text and metatext, (b) quality of the captured text and (c) compliance of the metatext to an established terminology. For the latter I use TEI markup on basically two levels: (1) markup for a precise identification of the attestation of the text and its specific editorial features and (2) markup to enrich the text from the perspective of my own research interests.

The presentation will provide examples for the applied markup. I will argue that the use of tag sets within the first category is certainly an indispensable prerequisite for long term efforts to build larger and larger corpora of Indic texts. The tag sets within the secondary category, on the other hand, seem to be of no relevance for this goal. The energy invested in the refinement of technically demanding tag sets is an asset for scholars who are so inclined. In the current state of Digital Indology, however, it is still necessary to develop standards for the discipline as a whole before we can start to agree on the basic ones.

Encoding history in TEI: A corpus-oriented approach for investigating Tibetan historiography

Mathias Fermer (mathias.fermer@oeaw.ac.at)

My presentation addresses a system for deriving historical evidence from Tibetan primary sources by applying semantic markup to the texts' key entities (i.e. persons, places, literary works and artefacts). This markup system follows the TEI-P5 guidelines and has been developed in the framework of the Sakya Research web-application (https://sakyaresearch.org/) which holds a large corpus of machine-readable sources in Classical Tibetan, ranging from medieval chronologies and histories to illustrious life stories of Buddhist masters.

The markup applied to the digital collection has been designed in line with the historiographical nature of the texts: It captures information about historic agents, the places they visited, as well as artefacts and literary works mentioned in varying contexts along the chronological sequence of the individual texts.

Using TEI-markup in this way has proven particularly useful in my own research for depicting the social, geographical, artistic and doctrinal contexts of the texts' narrative subjects (and their authors). It allowed for tracking teacher-student relationships and exploring the geographic expanse of those masters' regional networks, to give two examples for how the empirical evidence from TEI-encoded texts can be assessed.

I will address the concept behind this markup and its potential for a quantitative, intertextual analysis that goes beyond single texts. What can Tibetan historians gain from markup-technology, if systematically applied to a wider corpus of literature?

I will argue that the data deriving from a consistent annotation of primary literature on a large scale will gradually change our understanding of Tibetan history. At the same time, such a corpus-oriented approach to historiography raises several conceptual and practical questions about how, and in which form encoded information can best be stored, analysed, displayed and reused.

Panels

Session Fri 3b: TEI and models of text III

Session abstract: Graphs - charters - recipes: challenges of modelling medieval genres with the TEI

Medieval studies had and still has a substantial influence on the TEI: medievalists actively participate in the development and revision of the TEI guidelines and many medieval digital editing projects use the TEI (e.g. Patrick Sahle lists 144 medieval editions in his catalogue and most use TEI). The TEI is flexible and expressive when it comes to modelling features of the most common types of medieval texts such as prose and lyrical works. However, new editing projects and less-typical sources challenge the standard and raise the question of whether the TEI can (or should) be adopted to accommodate the special requirements of different genres or whether other standards should be used in conjunction with the TEI.
Using three medieval genres as case studies, this session will discuss the challenges encountered when trying to model non-literary texts (diagrammatic stemmata, charters, and recipes) and present solutions for how they may be modelled in both medieval and modern contexts.

Paper 1: Text - Graph - Image: Towards a Digital Edition of Peter of Poitiers' Compendium historiae

Authors: Roman Bleier, Franz Fischer, Tessa Gengnagel, Patrick Sahle, Andrea Worm

Abstract

Around 1180, Peter of Poitiers (c. 1130-1205), theologian at the cathedral school in Paris, compiled a survey of biblical history in the form of a diagrammatic stemma. The *Compendium historiae* presents the history of salvation in its linear and teleological structure. Biblical books and stories are condensed to biographical information and arranged in a strict chronological order. History, as it were, is mapped out in the form of a graph, thus allowing readers to assess synchronicity and diachronicity of people and events at a glance. Since this was deemed highly useful, Peter of Poitiers' work soon became the template for other chronicles that used a similar diagrammatic format to present history.
Despite its popularity and enormous impact, a scholarly edition of the *Compendium historiae* remains a desideratum to this day. It is primarily the graphical nature of the text that has prevented such an edition (or even for the work to appear in print). Another problem results from the structural and graphical variance among the manuscript witnesses and from the difficulty in representing graphical variants in the constituted text or critical apparatus of a (printed) scholarly edition.
The project team would like to use the opportunity of the TEI conference to present and discuss a preliminary data model for the representation of the *Compendium historiae* and the complex composition of its textual, pictorial and graphical elements. In that context, for instance, the TEI module "Graphs, Networks, and Trees" will be discussed as a means to represent the genealogical and historical lines of the *Compendium historiae*. The modelling

and presentation of a work that consists of a graph as well as adjacent texts, diagrams and images will provide a template for a great number of medieval and modern works that use graphic visual means to convey information.

Suggested reading

Worm, Andrea. 2013. “Visualising the Order of History: Hugh of Saint Victors' Chronicon and Peter of Poiters' Compendium Historiae.” In *Romanesque and the Past: Retrospection in the Art and Architecture of Romanesque Europe*, edited by Richard Plant and John McNeill. Leeds: Maney.
Worm, Andrea. 2019. *Geschichte und Weltordnung: Graphische Modelle von Zeit und Raum in Universalchroniken vor 1500*. Berlin: Deutscher Verlag für Kunstwissenschaft.
TEI. “19 Graphs, Networks, and Trees.” P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.6.0. Last updated on 16th July 2019, revision daa3cc0b9. Accessed July 19, 2019. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/GD.html.

Paper 2: Modelling charters in TEI P5: the TEI_CEI ODD

Author: Sean M. Winslow

Abstract

Charters and other documentary materials have their own traditional disciplinary approaches and concerns which are related to manuscript materials, but incompletely modelled by the msdescription module. A specialist fork of an older version of the TEI, the Charters Encoding Initiative, exists and is implemented for over 600,000 items in the Monasterium.net portal, but is showing its age as the TEI has received more updates which are not fully implemented in the CEI fork. Additionally, the non-standard nature of the CEI means that the training and resources for learning and extending the TEI are incompletely available to the CEI ecosystem. As part of the FWF-funded project ‘Retain Domain Specific Functionalities in a Generic Repository with Humanities Data” (ORD84)” (PI: Georg Vogeler),’ the CEI has been updated to be compatible with the TEI P5, including an ODD extending current manuscript description elements to include specialized diplomatics material. This talk will present the ODD, discuss it in the context of how the needs of modelling documentary materials go beyond the currently available TEI elements, and outline current efforts to update the large database of digitized charters and charter records to be compatible with the most recent version of the TEI.

Suggested reading

Ambrosio, Antonella, Sébastien Barret and Georg Vogeler, eds. 2014. *Digital diplomatics The computer as a tool for the diplomatist?* Beihefte zum Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde 14. Böhlau.
Vogeler, Georg, et al. 2018. “Charters Encoding Initiative.” http://cei.lmu.de

Paper 3: Starving for TEI? Cooking Recipes of the Middle Ages - Corpus, Analysis, Visualisation

Authors: Helmut Klug, Christian Steiner, Elisabeth Raunig

Abstract

CoReMA puts an interdisciplinary focus on the cross-cultural research of medieval cooking recipes and their interrelation. The project is preparing the cooking recipe transmission of France and the German speaking countries, which sums up more than 80 manuscripts and about 8000 recipes, for the analysis of their origin, their relation, and their migration through Europe.

The TEI model for the medieval sources is based on their page structures, while the transcribed text is annotated deeply, reaching below character level using the TEI gaiji-module. Optical and haptic descriptions of the manuscript including measurements, description of the writing material, order of quires and a description of the manuscript's content are encoded within the TEI's manuscript description. The philological level of the project can thus be covered thoroughly and effectively with the TEI.

However, the challenges of CoReMa involve a systematical semantic annotation of the cooking recipes' contents like ingredients, preparation instructions and time, tools, serving suggestions and medicinal, cultural as well as religious implications in the texts. The sheer quantity of semantics within one historic recipe shows how complex a semantic annotation of these texts is. The TEI does not sufficiently cover the needs for a comprehensive semantic encoding for cooking recipes. This is why we chose to develop our own annotation schema and implement it with an encapsulating TEI schema, taking on the TEI whenever reasonable. We will subsequently use RDF as the base format for exchange and analysis giving us the ability to work outside the barriers of historical and language constraints. The questions raised for the TEI is to consider how such an implementation of one's own needs can succeed without losing the advantages and ideas of the TEI.

Suggested reading

Michigan State University Libraries Special Collections, n.d. “Feeding America: The Historic American Cookbook Dataset.” Accessed July 28, 2019. https://www.lib.msu.edu/feedingamericadata.

Honkapohja, Alpo. 2013. “Manuscript abbreviations in Latin and English. History, typologies and how to tackle them in encoding.” *Studies in Variation, Contacts and Change in English* 14, ch. 2.4. http://www.helsinki.fi/varieng/series/volumes/14/honkapohja/.

Stokes, Peter A. 14.10.2011. “Describing Handwriting, Part I-VII: Recapitulation and Formal Model.” Blog, DigiPal. Accessed July 28, 2019. http://www.digipal.eu/blog.

TEI for manuscript description: progress, problems, and potential

The TEI module for manuscript description has been widely adopted since its creation. It is currently used by a number of major cataloguing and digitization projects, dealing with manuscripts from the Western, Greek, Slavonic, Islamicate, Syriac and Ethopian traditions.

The widespread adoption of TEI for manuscript description should be a cause of celebration for the TEI community and for manuscript scholars alike. The TEI standard allows for highly granular encoding, enabling complex research questions to be answered, and these questions can be pursued across an increasingly wide corpus of material. Scholarly hopes for a union manuscript catalogue and a large corpus of material for quantitative and comparative codicology are closer to being realized.

At the same time, the widespread use of the module should stimulate reflection and, ideally, act as a spur to increased collaboration. How far does the module currently meet the needs of cataloguers and scholars from different manuscript traditions? What priorities for future development can be identified? What are the divergences between implementations of the module in different projects, and how far do they enable or impede interoperability?

The aim of this panel is to provoke discussion and reflection on the progress that has been made, the problems that remain outstanding, and the potential for further development. The panel will comprise four short papers of 15 minutes each, representing cataloguers and researchers from different manuscript traditions. They will be followed by a 5 minute response from a named respondent, with 25 minutes for discussion.

Papers:

1. Report on Symposium on Manuscript Description – Yasmin Faghihi and Huw Jones

We have been working with TEI for manuscript description for nearly ten years, creating large datasets and complex descriptions for catalogues and digital libraries. While we have benefitted greatly in terms of making individual collections or projects available online, we are unsure if we have fulfilled the potential which TEI offers for the creation of truly interoperable data, for shared development around standard datasets, and in the use of that data by researchers. In each of these areas, we feel that expectations are not necessarily matched by day-to-day experience.

Efforts in community building around interfaces like the FIHRIST union catalogue have been largely successful. In the wider context, however, we have often encountered an initial resistance towards TEI. While it is recognised as a cutting-edge standard, academics, developers and curators often argue against the use of TEI owing to its complexity and the lack of standardised inputting tools.

We organised a symposium to bring together librarians, developers and researchers to investigate these issues and discrepancies. What do these diverse groups really think of TEI as a standard for manuscript description? Why are some projects and institutions using it and not others? For those that are using it, what is their experience of creating interoperable data in TEI? And what is the experience of a researcher using this data to answer research questions?

In this paper we will discuss the context of the Symposium and give a brief overview of the content of each session. But our main focus will be on the outcomes – what emerged which was useful or surprising, and how we might turn ideas and issues arising from the event into concrete actions.

Biographies

Yasmin Faghihi is Head of the Near and Middle Eastern Department of Cambridge University Library and in charge of a collection of over 4500 Islamic manuscripts. She has been involved in the creation of FIHRIST the online union catalogue for manuscripts from the Islamicate world and in the implementation of TEI and best practice for manuscript description since 2009.

Huw Jones is Head of the Digital Library Unit and Digital Humanities Coordinator at Cambridge University Library. His research interests include open technologies and standards for digital humanities including manuscript description and transcription in TEI and automated image analysis using IIIF.

Keywords

manuscripts; symposium; development; libraries; research

2. A Style Guide for Western Manuscript Description: Common Practice and Future Steps - James Freeman and Matthew Holford

This paper presents the results of a two-year collaboration between Cambridge and Oxford to produce enriched, shared guidelines for the description of western medieval manuscripts in TEI. This 'style guide' facilitates the use of TEI by cataloguers by mediating the P5 guidelines through plain-language, example-based explanations of encoding, in harmony with existing descriptive practices. It also increases the value to researchers of the metadata being generated in institutional libraries, enabling the production of consistent and comparable descriptions, by establishing controlled vocabularies and providing detailed examples of encoding from more complex manuscripts. In consultation with researchers we have developed an encoding framework which can capture information for the key research questions of western codicology.

Compiling the style guide has forced us to confront many practical and intellectual questions. We have refined the P5 guidelines within the existing structure, or made concrete suggestions for its revision. We cover many issues treated briefly or not at all in the TEI Guidelines, including guidance on glossed manuscripts, pseudonymous authors, added texts and decoration, detailed description of paper, and the treatment of collation structures. However, there are also outstanding issues for which no obvious solution has yet presented itself. This conference represents the perfect opportunity for us to engage with and invite feedback from the wider TEI community. Our ambition is to make the style guide freely and publicly available, to broaden the scope of the collaboration, and to include libraries, cataloguers, researchers and the TEI community in further, ongoing dialogue about its future development and implementation.

Biographies:

Dr James Freeman, Medieval Manuscripts Specialist, Cambridge University Library
Dr Matthew Holford, Tolkien Curator of Medieval Manuscripts, Bodleian Library, University of Oxford

Keywords

manuscripts, medieval, codicology, metadata, libraries

3. Theory and practice of Using TEI for manuscript cataloguing – Torsten Schaßan

In Germany, there's a long tradition in (extensive) manuscript cataloguing. While the MASTER format, and its successor, the msDesc module of the TEI, have been very much oriented at the cataloguing practice of various cataloguing traditions, there are still a lot of challenges in encoding and processing manuscript descriptions in TEI. Basic issues are:

- importance of visualisation ("What would the text look like in print?")
- importance of whitespace and punctuation
- paragraph structures (e.g. layout, decoNote, provenance, etc) vs. elementary information

Another topic is the question of how electronic descriptions become available?

- What are the differences between conversion of a printed catalogue to a digital text and born-digital descriptions?
- Is there a difference in the encoding of texts which are meant for print and a basic full text search than of texts optimised for detailed, and expert search?

Finally, there's the question of which information in which fields is needed at a minimum to find a certain manuscript? In Germany, in the course of the development of a national manuscript portal, a set of core fields has been defined which could probably be used to define a level of interoperability of manuscript description data.

Biography:

Torsten Schaßan works in the Manuscript and Special Collections Department of the Herzog August Bibliothek Wolfenbüttel. He has been active in multiple projects on cultural heritage digitisation and cataloguing. He is an active member of the TEI community and member of the TEI-MS-SIG. His research interests are digital humanities, digital editions, text encoding, and markup theory.

Keywords

cataloguing; medieval manuscripts; Handschriftenportal; core fields;
production, presentation, and use of TEI documents

4. Medieval Manuscripts in Oxford Libraries and Manuscripts from the Islamicate World: data, metadata, interface, and their role in research - Dr Luca Guariento

The release of TEI-based catalogues such as the "Western manuscripts at the Bodleian Libraries and selected Oxford colleges" and of the "Union Catalogue of Manuscripts from the Islamicate World" (FIHRST) is warmly to be welcomed. In particular, their availability on GitHub is a forward-looking way of sharing this data. Both of the web resources use the same technical infrastructure. How can this form of presentation aid research? The potential for quantification, for instance, is evident. Also, the flexibility of the TEI, combined with the user interface's faceting options, allows for some filtering and statistics.

Nevertheless, there is definitely some room for improvement. For instance, an important aspect of a manuscript is its dimensions: the web interface does not allow filtering or faceting according to this description. Therefore, questions like: “Were small manuscripts more likely to be decorated?” “Were large manuscripts more frequently produced in a specific century” “Did different countries have preferences for different sizes of manuscripts?” cannot be easily answered. Another example is musical notation, which is often concealed and difficult to identify: at the moment a filtering/faceting by 'music' is not possible.

Exploring the same data via GitHub provides more complex ways of research and analysis, although of course this is not for the average user, and everybody likes a nice and friendly interface. Does the web interface play a role in facilitating (or obstructing) research in these catalogues? Or is good research only dependent on the TEI schema, the data quality, and its consistency?

Biography

Dr Luca Guariento completed his PhD in Music at the University of Glasgow, researching on the seventeenth-century polymath Robert Fludd. In the meantime, he got increasingly involved in many Digital Humanities projects, specialising in XML technologies and TEI standards, IIIF, and other DH things. He is now Research Systems Developer at Glasgow University, where he combines his grounding in the Arts and Humanities with his passion for digital humanities tools and methodologies.

Keywords

manuscripts; analysis; quantification; distant reading

Document Modeling with the TEI Critical Apparatus

A Panel Proposal

Overall Abstract

This panel addresses the TEI critical apparatus as a data model, investigating how it has expanded the capacity of scholarly editions to articulate and analyze phenomena of textual variation and multiplicity. We will discuss how the TEI critical apparatus, as a structure that mediates distinct versions of a work, is expanding horizons for multidimensional and pluralistic document modeling. Our panel surveys recent experiments with the critical apparatus that have led to new kinds of scholarly research and in some cases to revisions to the TEI Guidelines. What kinds of research questions and applications can we support with the TEI critical apparatus, and what practical challenges do we face in working with it in inline and stand-off ways?

We begin by investigating how the TEI critical apparatus has transformed the expressive capacity of scholarly editions to prioritize textual multiplicity. We continue by sharing data models that apply TEI critical apparatus as a stand-off “spine” for connecting independently encoded witnesses. We conclude by inviting the audience to discuss with us the scalability of these methods for texts with large numbers of witnesses, and the technological challenges and opportunities of stand-off methods in light of recent changes to the TEI Guidelines.

What is a Critical Apparatus, Really?
Hugh Cayless

A critical apparatus in a printed scholarly edition is the set of notes made by an editor in support of their presentation of a text. The TEI Guidelines’ Critical Apparatus module builds on the notion of the critical apparatus in print to do something different and more powerful—to actually model textual variation rather than simply note where it occurs. Textual variation and its expression is particularly complex. Texts can vary along a number of axes: an editor may be faced with a work that has a number of similar versions, so that it may be possible to infer an original text; more significant variation may happen, so that while a single "work" is identifiable, no single text exemplifies that work; several differing stages in the development of a final edition may be extant; a single document may have multiple, differing editions. There are also different philosophies of edition-making that vary across discipline and document type. This presentation will set the stage for the panel by defining the intellectual space occupied by different types of Digital Scholarly Edition and providing examples of the ways different kinds of DSEs are realized. A number of these efforts are predicated on the idea that the critical apparatus is obsolete. What then does that mean for the Critical Apparatus module?

This is (not) Spinal Tap: Modeling to Prioritize Variance
Elisa Beshero-Bondar

This and the following paper are related, based on the development of a stand-off “spine” for variorum editions. Working on the *Frankenstein Variorum* project, we designed a model that we call a “spine” as a “backbone” or “binding,” designed to prioritize the reading of changes to a work over time rather than to marginalize such reading in the tradition of the print apparatus. The “spine” begins as the edited output of machine-assisted collation, and is transformed into a stand-off document containing pointers and data about variant passages. It is used to generate reading views of each edition that can either:

a) Generate TEI XML editions of each distinct version, or
b) Highlight passages in an existing TEI edition without altering that edition at all using TEI XPointer schemes.

In the first case, the apparatus can construct new TEI editions, locating moments of alteration based on alignment information stored in the “spine”. In the second, the apparatus alone is needed to connect with existing editions. Through the “spine” we can pull something as complex as TEI page-by-page manuscript encoding—without altering it—into a new web interface that points out where the manuscript semantically aligns with and varies from the printed editions. Thanks to the document model of our “spine” we produce from collation data a lightweight and readily updatable mechanism to bring multiple kinds of editions into comparative view.

Publishing a stand-off critical apparatus: Leveraging isomorphic representations across text and music notation
Raffaele Viglianti

The “spine” collation described in the previous paper relies on stand-off markup techniques and pointers in order to represent variance across a number of sources. In building a web based publication out of this structure, the code will need to follow these pointers to locate, collect, and publish these resources through an interactive website. This would be quite a challenge with traditional transformations of TEI data to HTML structures, because the pointers would need to be reinterpreted to locate data in the HTML surrogate. CETEIcean, a tool for rendering TEI directly in the browser, creates an isomorphic representation of a TEI document as HTML Custom Elements. This direct correspondence between the TEI data and its surrogate makes it easier to follow pointers from the “spine” collation and build interactive publications. This presentation will introduce the Early Modern Songscapes project, in which this approach has been tested for both text and music notation. It is possible to apply this approach to music notation data because the Music Encoding Initiative format (MEI) provides structures equivalent to the TEI critical apparatus and the tool Verovio generates an isomorphic SVG representation of MEI data to be published as an interactive score in the browser.

Response: Data models, many-witness texts, and the future of apparatus markup
James Cummings

In the final contribution James Cummings will respond to the previous three panelists’ presentations, by reflecting on their views of the TEI critical apparatus markup as a data model that both expresses interpretations of multi-witness texts and creates structures for comparative analysis. He will respond to the practical challenges the previous panelists have described with open-ended questions designed to lead into discussion with the audience. He will call attention to the difficulties and lack of technological support for individual projects using TEI critical apparatus markup through stand-off or out-of-line methodologies. He will question how scalable these solutions might prove to be for those working with traditions that involve a significant number of witnesses, or whether there may be similarly progressive workflows for encoding copy-specific apparatus for large witness groups. Part of the response will highlight for the audience some of the more recent modifications to the TEI Guidelines in this area, for example the move from allowing just phrase-level content in <rdg> elements to larger structures such as <div> and <floatingText>, and the additional challenges this may introduce. In doing so, this response will question where TEI critical apparatus markup might develop in the future.

Posters

A TEI-based model to encode notarial charters (Asturias, 1260-1350 ca.)

Elena Albarrán-Fernández

The transformation of social, political and administrative models all along the Mediterranean medieval cultures–Italy, French Midi, Christian Iberia–, made notaries public a powerful social agent. As part of a new cultural and laic social group, notaries acted as a bureaucratic link between the common people and the elites, allowing the access not only to writing, but also to legally valid instruments that would resist memory. Its nomination was disputed between the different authorities of the Castilian Crown –monarchy, Church and nobility–as an element of jurisdictional power. The study is focused on the notarial institution and its documentary production from mid XIIIth to mid XIVth centuries. Most of these documents come from Benedictine and Cistercian monasteries established in Asturias–San Pelayo and San Vicente of Oviedo or Santa María of Belmonte are some of them–, but also from Oviedo's Cathedral and some relevant city councils back in medieval times. Nowadays, many of the documents that make up the corpus are still preserved in this northern region, while another important amount of them is stored at the Archivo Histórico Nacional, in Madrid. The dispersion of these documentary collections adds more difficulties to the task: in the best-case scenario, there is already a digitised version of the document–e.g. in Portal de Archivos Españoles (PARES) , also accessible at Monasterium (MOM) –; if the first option is not available, the researchers have to go where the documents are preserved to proceed with their work. The foundations of this research project were established in 1986 at the VII Congress of the Commission Internationale de Diplomatique (Sanz, 1989). From this moment on, studies about notarial documentary production and notarial institution have evolved. Research projects focusing on early stage notarial history share three common purposes: to trace the different phases in the implantation of a renewed institution, and to measure the gradual reception of the restored Roman Law and specific legislative works; to observe the transformation of documentary forms and check the increasing use of writing and written instruments; and to document the jurisdictional and political tensions for the control of notaries public. Nevertheless, the methodological approaches have evolved since then according to the needs of recent research projects. Larger documentary corpora, massive digital repositories or complex crossed queries are some of our current assets to push diplomatics–and digital diplomatics– research further (De Paermentier, 2014). To achieve those objectives, the methodology developed attended to several aspects: on a historic-diplomatic basis, we have applied a TEI-based model–following the example led by the Charters Encoding Initiative (CEI) – attending to archival data, diplomatic features and socio-political aspects. Each manuscript has its own XML document, as the corpus is constructed with approximately 400 individual documents. The three main elements of our TEI-based model are: fileDesc, contains the information related to our thesis project and research group and the archival file sheet of each document; profileDesc, contains the abstract and the diplomatic data such as the actio, redactio, traditio or the datatio; and the body, contains the transcription of the document, where the element seg identifies each part of its diplomatic structure, using attributes such as @function, @type and @subtype. This mark-up model was conceived to apply a systematic analysis of the data, in order to study the evolution of notarial documents in this early stage period. For example, the use and progressive loss of religious content in the formularies used by these first notaries public. Religious formulas can be found in the invocatio and spiritual penal clauses so, using XPath queries we have managed to: trace a chronological evolution of the invocatio in our corpus, registering the maintenance and progressive loss of this kind of formulas through a century– strong presence of invocatio formulas between 1270 and 1320 –. We have noted nine different variants of the invocatio

-from its latin form In nomine Domini, amen to a more evolved romance variant En el nomne de Dios e de la Virgen Santa María sua madre, amen-; from the 142 documents that contain an invocatio formula, we have identified the most recurrent juridical actions that make use of it– mostly sales, donations and avecindamientos (i.e., a neighborhood charter) –. In this work, we have exposed a mark-up system (based on TEI and CEI) which we are currently applying to extract, on a systematic basis, data from a wide range of documentary typologies. Furthermore, it can be used on other kind of documents such as the ones produced by pontifical and royal chancelleries. As future work, a more complex strategy of data analysis will be designed and applied. Our main purpose is to build a complete characterisation of the diplomatic tenor of the documents contained in our corpus. It will serve as an experimental database to test different types of analysis on large-size corpora and to produce a long-lasting encoding model.

An Exploration of <object> using Antarctic Artifacts

The recent inclusion of <object> to the TEI has been a long-awaited addition to the language. The ticket proposing the creation of an <object> element (https://github.com/TEIC/TEI/issues/327) outlines some of the discussions that have led to this implementation. However, as with all TEI elements, there is room for expansion and refinements as new use cases are introduced. In this poster, we will explore the usage of <physDesc> element within <object> to gauge how this adaptation of the msdescription module meets the needs of encoding more general objects.

We will evaluate <physDesc> in this new context by creating examples of the usage of <object> by describing several items from the Robert E. Hancock Jr Antarctic Collection at FSU Special Collections (https://archives.lib.fsu.edu/repositories/4/resources/642). We have selected a diverse set of objects from this collection, including a model Navy Icebreaker and a commemorative plaque, to cover a wide variety of use cases. The objects will be described in as much detail as possible using the existing <physDesc> with its child elements. Our methods have two outcomes: to determine whether the current adaptation of <physDesc> for <object> is appropriate for these objects and to produce more examples of <object> to be included in the Guidelines.

In order to expand the utility of the <physDesc> element, we will implement an ODD customization to propose ways of identifying subsections of a given object (e.g. the base of a figurine as distinct from its top) using <objectDesc>. Our ODD customization will include a new element that will allow us to differentiate between descriptions of different sections of the object and provide two mechanisms for disambiguating pieces of the object: one more generic and a coordinate space version based on the att.coordinated attributes.

Word Count: 300/300

Promoting Bellini’s legacy and the Italian opera by scholarly digital editing his own correspondence

Angelo Mario Del Grosso[1], Erica Capizzi[2], Salvatore Cristofaro[3], Graziella Seminara[4], Daria Spampinato[3]

[1] Istituto di Linguistica Computazionale - CNR, Italia - angelo.delgrosso@ilc.cnr.it
[2] Università di Catania, Italia - ericacapizzi@yahoo.it
[3] Istituto di Scienze e Tecnologie della Cognizione - CNR, Italia - salvatore.cristofaro@istc.cnr.it, daria.spampinato@cnr.it
[4] Università di Catania, Italia - g.seminara@unict.it

KEYWORDS

Digital Scholarly Edition, Correspondence, Digital and Computational Philology, Vincenzo Bellini, Music

ABSTRACT

This contribution aims at illustrating the ongoing work towards the digital scholarly editing, long-term preservation, web publishing and computational exploiting of 41 letters, written by the renowned composer Vincenzo Bellini. The correspondence is kept at the Belliniano Civic Museum of Catania and it is being encoded in XML according to the last TEI guidelines. The edition will be made accessible both via web - exploiting the Edition Visualization Technology (EVT)[1] (Rosselli Del Turco and Di Pietro, 2016) - as well as integrated into an interactive and multimedia tour within the museum (Del Grosso et al., 2018). The digital edition is based on the recently published transcriptions made by Seminara (Bellini, 2017). The encoding scheme has been defined according to the edition requirements (Bozzi, 2014), the TEI best practices (Pierazzo, 2015) and the Music Encoding Initiative (MEI) guidelines - where the musical context must be specified (Figure 1).[2] Our initiative has some elements of innovation that distinguish it from similar projects, such as the Van Gogh letter project[3] or the DALF project[4]. For instance, we encode the circumstance that the letters themselves have also the purpose of acting as envelopes (Figure 2). In fact, they are folded on themselves and postmarks and wax seals are sometimes affixed on them. The edition takes care of handling the correspondence metadata by means of the *correspDesc* TEI tagset[5] (Stadler et al., 2016), thus providing the opportunity to exploit the *correspSearch* API[6] (Neuber, 2016). This approach has allowed us to enrich the encoding of the document both in its logical and physical structure and in indexing letters by sender, recipient, date, and places (Figure 3).

The museum context and the educational purposes have even led us to the definitions of some lists of *named entities*. Within these resources we have adopted the Semantic Web and LOD paradigm by encoding external references to authoritative repositories such as RISM[7] and DBpedia.

Finally, we implemented some useful EVT extensions to automatically handle hotspots and to show critical notes that accompany the text[8] (Figure 4).

[1] http://evt.labcd.unipi.it/

[2] No letters preserved in the Belliniano Civic Museum contain fragments of musical compositions. MEI guidelines have only been exploited to encode the person role in the musical context (librettist, *tenore*, etc.) which occurs in the text.

[3] http://vangoghletters.org/

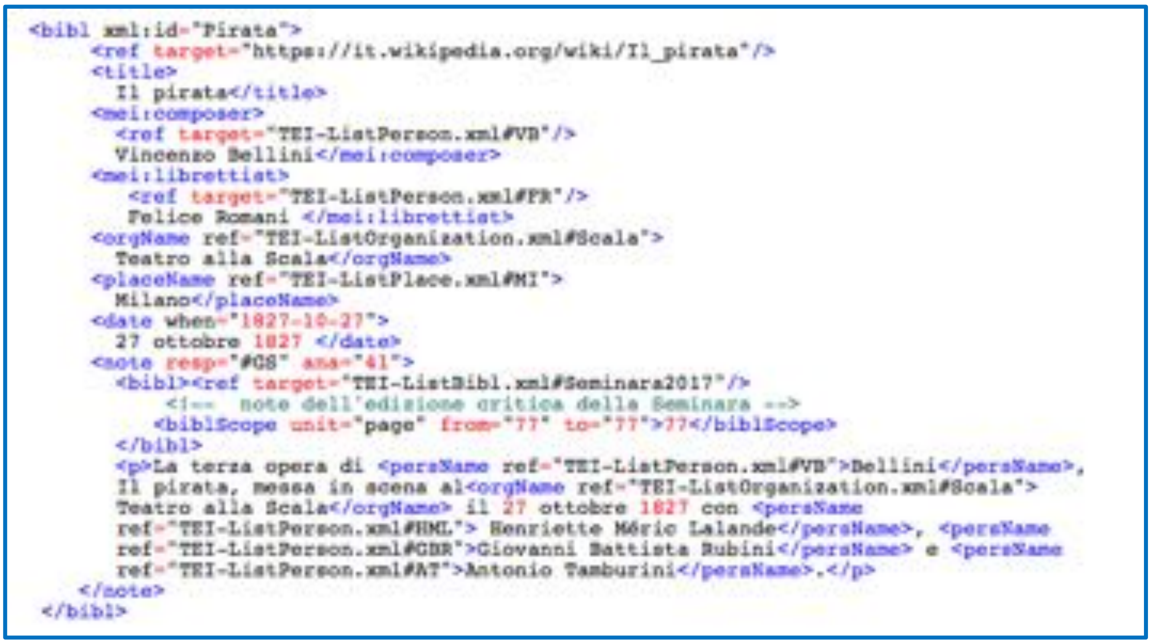
[4] http://ctb.kantl.be/project/dalf/

[5] http://www.tei-c.org/release/doc/tei-p5-doc/it/html/ref-correspDesc.html

[6] https://correspsearch.net/index.xql?id=api&l=en

[7] https://opac.rism.info/index.php?id=8&L=1#c116/

[8] http://licodemo.ilc.cnr.it/bellini-in-rete



```
<bibl xml:id="Pirata">
     <ref target="https://it.wikipedia.org/wiki/Il_pirata"/>
     <title>
       Il pirata</title>
     <mei:composer>
       <ref target="TEI-ListPerson.xml#VB"/>
       Vincenzo Bellini</mei:composer>
     <mei:librettist>
        <ref target="TEI-ListPerson.xml#FR"/>
        Felice Romani </mei:librettist>
     <orgName ref="TEI-ListOrganization.xml#Scala">
       Teatro alla Scala</orgName>
     <placeName ref="TEI-ListPlace.xml#MI">
       Milano</placeName>
     <date when="1827-10-27">
       27 ottobre 1827 </date>
     <note resp="#GS" ana="41">
       <bibl><ref target="TEI-ListBibl.xml#Seminara2017"/>
           <!--  note dell'edizione critica della Seminara -->
          <biblScope unit="page" from="77" to="77">77</biblScope>
       </bibl>
       <p>La terza opera di <persName ref="TEI-ListPerson.xml#VB">Bellini</persName>,
       Il pirata, messa in scena al<orgName ref="TEI-ListOrganization.xml#Scala">
       Teatro alla Scala</orgName> il 27 ottobre 1827 con <persName
       ref="TEI-ListPerson.xml#HML"> Henriette Méric Lalande</persName>, <persName
       ref="TEI-ListPerson.xml#GBR">Giovanni Battista Rubini</persName> e <persName
       ref="TEI-ListPerson.xml#AT">Antonio Tamburini</persName>.</p>
    </note>
 </bibl>
```

Figure 1. Bibliographic entry for Pirata Opera with MEI elements encoding the composer and librettist.

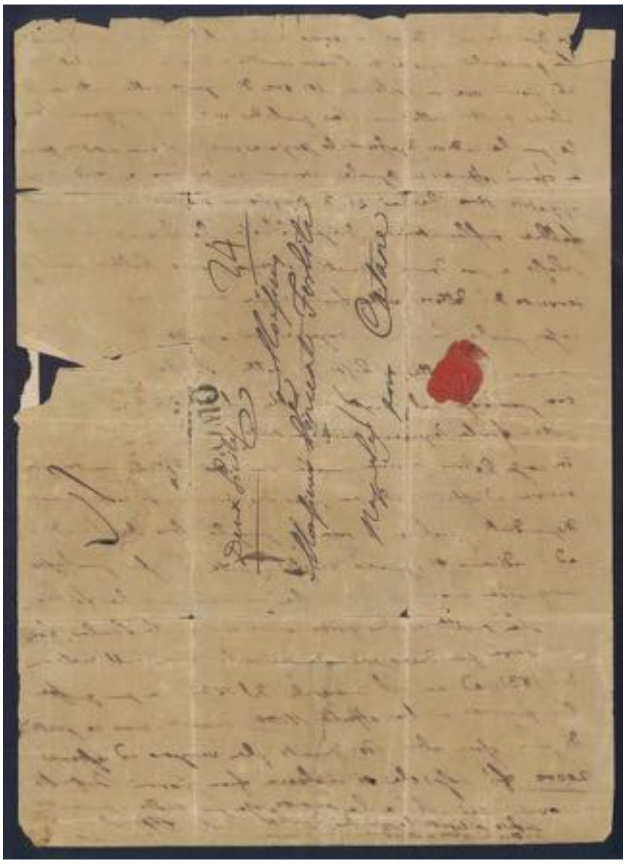
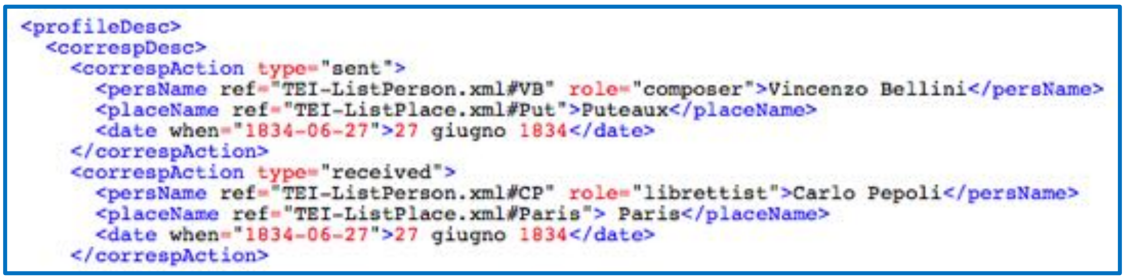
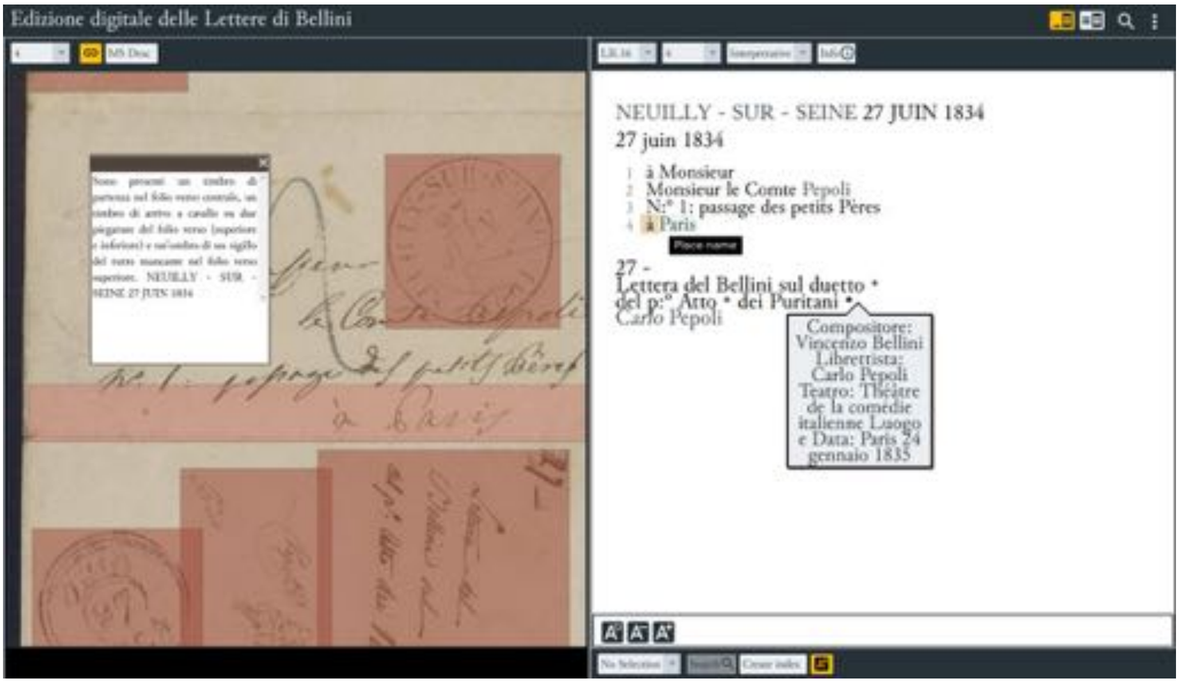


Figure 2. Enveloping of letter n. LL1.4.



```
<profileDesc>
  <correspDesc>
    <correspAction type="sent">
      <persName ref="TEI-ListPerson.xml#VB" role="composer">Vincenzo Bellini</persName>
      <placeName ref="TEI-ListPlace.xml#Put">Puteaux</placeName>
      <date when="1834-06-27">27 giugno 1834</date>
    </correspAction>
    <correspAction type="received">
      <persName ref="TEI-ListPerson.xml#CP" role="librettist">Carlo Pepoli</persName>
      <placeName ref="TEI-ListPlace.xml#Paris"> Paris</placeName>
      <date when="1834-06-27">27 giugno 1834</date>
    </correspAction>
```

Figure 3. The correspDesc tagset adopted to encode LL1.16 letter.



Edizione digitale delle Lettere di Bellini

Sono presenti un timbro di partenza sul folio verso centrale, un timbro di arrivo a cavallo su due piegature del folio verso (superiore e inferiore) e sull'ombra di un sigillo del recto mancante sul folio verso superiore. NEUILLY - SUR - SEINE 27 JUIN 1834

NEUILLY - SUR - SEINE 27 JUIN 1834
27 juin 1834

à Monsieur
Monsieur le Comte Pepoli
N:° 1: passage des petits Pères
à Paris

27 -
Lettera del Bellini sul duetto •
del p:° Atto • dei Puritani •
Carlo Pepoli

Compositore: Vincenzo Bellini Librettista: Carlo Pepoli Teatro: Théâtre de la comédie italienne Luogo e Data: Paris 24 gennaio 1835

Figure 4. Image hotspots and text notes.

ACKNOWLEDGMENTS

This work was partially supported by the "Patto per Catania" under the "Fondo Sviluppo e Coesione 2014-2020: Piano per il Mezzogiorno".

REFERENCES

Bellini, V., 2017. Carteggio. Leo S. Olschki.

Bozzi, A., 2014. Computer-assisted Scholarly Editing of Manuscript Sources, in: Davidhazi, P. (Ed.), New Publication Cultures in the Humanities: Exploring the Paradigm Shift. Amsterdam University Press, Amsterdam, pp. 99–115.

Del Grosso, A.M., Spampinato, D., Cristofaro, S., De Luca, M.R., Giovannetti, E., Marchi, S., Seminara, G., 2018. Le lettere di Bellini: dalla Carta al Web, in: AIUCD 2018 - Book of abstracts. pp. 60–64. https://doi.org/10.6092/unibo/amsacta/5997

Neuber, F., 2016. correspSearch, in: Variants. European Society for Textual Scholarship, pp. 284–285.

Pierazzo, E., 2015. Digital Scholarly Editing: Theories, Models and Methods, Digital Research in the Arts and Humanities. Ashgate, Farnham Surrey.

Rosselli Del Turco, R., Di Pietro, C., 2016. Between innovation and conservation: the narrow path of UI design for the DSE, in: Digital Scholarly Editions as Interfaces. Presented at the Digital Scholarly Editions as Interfaces, University of Graz, Graz, pp. 133–163.

Stadler, P., Illetschko, M., Seifert, S., 2016. Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>. Journal of the Text Encoding Initiative 9. https://doi.org/10.4000/jtei.1742

Annotating Cooking Recipes of the Middle Ages for semantic analysis and visualisation

The TEI mainly provides structural elements. Semantic annotation is primarily seen in names, dates and places and in metadata modules as, for example, the msDesc. Most of the projects include semantic information in attribute values. However, this approach generates a greater workload, annotated text that is more difficult to read for humans and most notably the interchangeability is reduced.

The analysis of our material will be based on Semantic Web technologies. By using concepts in the sense of a notion, an idea rather than a term, we are trying to overcome historical and language constraints. Once the entities of each recipe are equipped with concepts, the project's analysis can reveal concurring or deviating eating habits, text migration as well as the influence of neighboring countries on their respective cuisine. The vast implementation of ontologies in the natural sciences allows us to establish connections from historical eating habits to modern concepts of food and generate new knowledge for the domain of food history. The research data will also be the basis for spatial and temporal visualization and statistical evaluation.

However, to do so we developed our own schema for the systematical semantic annotation of cooking recipes' contents like ingredients, preparation instructions and time, tools, serving suggestions and medicinal, cultural as well as religious implications in the texts. Nonetheless, our goal was to use TEI elements wherever possible. Introducing semantically charged elements improved the workflow considerably and a mapping of these elements back to standard TEI could close the circle. It should be discussed, though, whether the use of a namespace for a foreign schema within a TEI capsule would provide a greater chance of exchangeability.

Denise Ardesi, Astrid Böhm, Helmut Klug, Bruno Laurioux, Corentin Poirier, Elisabeth Raunig, Christian Steiner

Abstract for a Poster at TEI 2019

Poster title: Encoding the documents of the “The Imperial Diet in Regensburg, 1576”.

Authors: Roman Bleier, Elisabeth Brantner, Gabriele Haug-Moritz, Christiane Neerfeld, Eva Ortlieb, Thomas Schreiber, Georg Vogeler, Florian Zeilinger

The digital scholarly editing project “Der Regensburger Reichstag von 1576” (The Imperial Diet in Regensburg, 1576) is a collaboration between the Historical Commission at the Bavarian Academy of Sciences and the Center for Information Modelling at the University of Graz. It is the latest edition in the long-term editorial endeavor “Reichstagsakten: Reichsversammlungen, 1556–1662” series and a pioneer project and will hopefully serve as a model for future editions of Imperial Diet records as it uses a “digital first” publishing approach.

The project has two primary goals: 1) to compile an overview over the rich written documentation of the event preserved in written records in various archives (several thousand documents) in the form of a list of manuscript descriptions; 2) to edit key documents of the “Reichstag” and enrich them using a TEI transcription model. A number of relevant documents have already been edited by the Historical Commission and exist in a simple text format. These texts will be converted to TEI and enriched with basic meta information using a semi-automated process. The focus of the project is on the many facets and complex communicative events at and around the Imperial Diet and the project team tries to systematically organise the information about these communicative events using an ontology. Using the TEI @ana attribute pointer to concepts in the ontology will be embedded in the TEI. It is hoped that this method will generate new knowledge about events, communication acts, participants, topics being discussed, etc. at and around the Imperial Diet of 1576.

The project team proposes a poster that will briefly outline the project and its goals, and present in detail the semi-automated digital enrichment workflow used to create the TEI transcriptions, the ontological model of communication will be presented and how it will be implemented in the TEI enriched transcriptions.

Virtual reconstruction of scattered provenance of Bohemian printed books

Authors: Renáta Modráková, Tereza Paličková

NAKI MK ČR project DG18P02OVV009 " Virtual reconstruction of scattered provenance of Bohemian books " 2018-2022 is handled by a consortium of three institutions (Library of the National Museum, National Library of the Czech Republic and National Heritage Institute). The main task of the project is to research and record Bohemian library collections (libraries and printed books by Czech owners and owner institutions) and their virtual reconstruction in the online database of the book owners PROVENIO and in the database of national authorities in the NL CR on the basis of Aleph (MARC21), which is supported by all XML-elements.The newly created authority records of the book owners will also be sent to the CERL Thesaurus database. The data is stored in the PROVENIO database. The project plans to upgrade the database with geolocations and to make it usable in GoogleMapps. The project results will be published through the database of book owners PROVENIO and in the form of five exhibitions (four standard and one virtual) dedicated to nobility libraries in Bohemia, Moravia and Silesia, surviving testimonies to the culture of reading in the past, nobility sponsorship of literature, and fragments of old library collections in the Reserve collection of the National Library. As a support for the care of historical collections in smaller museums, galleries and other memory institutions will be elaborated a methodology of research, revision, and passportisation of historical book collections and in cooperation with Czech Association of Museums and Galleries will be ready for experimental operation an on-line form for provence evidence.

Research of provenance glosses in medieval manuscripts and in incunabulas from historical collections of the National Library of the Czech Republic

Author: Renáta Modráková

The Department of Manuscripts and Early Printed Books of the National Library of the Czech Republic started in 2017 a systematic and comprehensive research of provenance glosses in medieval manuscripts and incunabulas deposited in the historical collections of the National Library. The primary goal of the project is to complete all medieval provenance glosses and to identify the private and institutional owners. First, data from available sources (i.e. all known and available printed catalogs and descriptions in digital databases and virtual libraries) are being reviewed, supplemented or/and corrected. The newly found glosses in manuscripts have been rewritten in accordance with the XML-structure used for virtual library of historical collections Manuscriptorium (in accordance with the TEI Guidelines P5). Glosses in incunabulas will be processed in Aleph platform (MARC21). In the last year of research, all completed records will be uploaded to the virtual library Manuscriptorium, hosted in the National Library CZ. The team has been continuously working on the identification of individual owners and ownership institutions, that will be shared with CERL Thesaurus. The ideal goal is to complete the original libraries, which were scattered in historical collections of the National Library (including geolocations). The reconstructed virtual medieval libraries will be accessible in the Manuscriptorium . This comprehensive research will allow further identification of significant cultural persons, geolocations, former libraries. This key research will complement existing scattered information about many original manuscripts and incunabulas and extend next research.

TTHUB: Text Technologies Hub for Extending TEI Training in Spanish

Susanna Allés Torrent (University of Miami)
Gimena del Rio Riande (IIBICRIT, CONICET)

The Text Encoding Initiative (TEI) was, from its beginnings, a very much Western, English-language project. However, its use is nowadays global. The efforts to internationalize the TEI's documentation started with an initiative led by Sebastian Rahtz by 2005. This work resulted in partial translations in Japanese, Chinese, Korean, Italian, Spanish, French, and German. Still, the TEI Guidelines are written and edited entirely in English, and most of the courses and tutorials are either written in English or use examples from European or North American literatures.

We believe that, in order to adopt the TEI, the academic community, individuals and institutions, need to have more resources available in their native language. With this in mind, we have recently decided to undertake a still emerging project- the Text Technologies Hub or TTHub[1]-, an online open site that aims to function as a hub of materials in Spanish devoted to the global Spanish community interested in Digital Scholarly Edition with TEI and Text Technologies. Our goal is to serve as a hub of available online materials, resources, software, and technologies that can potentially serve those Spanish-speakers scholars interested in textual studies, digital editing, corpus construction, and digitization processes in Spanish. It can also help researchers interested in Hispanic Literatures.

This poster aims to describe the goals, challenges, possible uses, and materials of the TTHUB. We believe resources as such can facilitate the learning experience of scholars and students in a self-taught experience, as part of the materials of an academic syllabus, or even as the common ground for teaching at workshops and other training events.

Susanna Allés-Torrent is Assistant Professor in the Department of Modern Languages and Literatures at the University of Miami, where she teaches Digital Humanities, Medieval and Early Modern Romance Literature.

Gimena del Rio Riande in Adjunct Researcher at the Instituto de Investigaciones Bibliográficas y Crítica Textual (IIBICRIT-CONICET, Argentina) and the Director of Humanidades CAICYT Lab at Centro Argentino de Información en Ciencia y Tecnología (CAICYT-CONICET). She is also External Professor at Universidad de Buenos Aires (UBA).

[1] See: http://tthub.io

Mary Erica Zimmer and Janelle Jenstad
2019 Text Encoding Initiative Conference
Graz, Austria
Poster Abstract (29 July 2019)

Documenting Discoveries: TEI and *Browsing the Bookshops in Paul's Cross Churchyard*

Traditional print scholarship mediates between the reader and the archival sources. We rely on the expert scholar to find, read, interpret, and digest sources for us. Yet even the most reputable scholars at times rely on searches that feel "hurried and incomplete" (Blayney, 1990, p. 1). By allowing us to present source documents, TEI-encoded transcriptions thereof, and scholarly conclusions based thereon, the digital environment affords the possibility of revisiting archival foundations. As early as 2000, Seamus Ross heralded the promise of "[d]igital archives combined with new technologies" to enable "simultaneous access to a range of sources" that would develop "research methods not possible with . . . printed or handwritten records" alone (Ross, 2000, p. 12). The typical response to this potential--exemplified by resources like *Shakespeare Documented*--has been to create new digital archives uniting disparate artifacts in digital space. *Browsing the Bookshops in Paul's Cross Churchyard* aims both to unify and to interrogate: revealing the document-based findings of prior projects while creating infrastructure able to support further research.

Building on the basis established by Peter W. M. Blayney's work allows his scholarly monuments to serve as literal foundations: ones to be excavated archivally while serving as points of departure for digital development. Our initial focus is his landmark 1990 study, *The Bookshops in Paul's Cross Churchyard*, which uses meticulous archival research to map bookshops and stalls near London's St. Paul's Cathedral during the period before the 1666 Great Fire. Among this volume's most influential contributions are its print visualizations: its composite, layered maps of the Churchyard, its shops, and their locations. At present, these maps are glossed through brief companion narratives conveying data points crucial to each rendering. Our work maximizes the affordances of digital media by connecting aspects of Blayney's reconstructions to documents that underpin them, as well as to related further texts and subsequent scholarship. Resulting will be an extensible TEI-based environment able to facilitate exploration of these intersecting document-based worlds.

Our work concentrates first on a single year of the Cross Yard's history (1600), coupled with in-depth analyses of five discrete shop sites (four from stationer Reyner Wolfe's "charnel chapel" group, plus his rental for The Brazen Serpent). Developing these two dimensions in tandem will establish guidelines for synchronic and diachronic renderings: those addressing the quantitative claims of intersecting datasets at a given moment, while grappling with challenges of gathering key spatial details over time. Both goals require interoperability: the latter, among XML schemas of multiple extant collections, including the *Map of Early Modern London* (MoEML), the *Stationers' Register Online* (SRO), and *British History Online* (BHO), to name a few. We are also exploring opportunities for connecting to related projects and corpora, including the 25,000+ texts of the Phase I EEBO-TCP corpus (with Phase II to follow). Ultimately, TEI's role as the project's *lingua franca* makes tangible the promise of interoperable research offered by humanities markup itself.

References

Blayney, P.W.M. (1990) *The Bookshops in Paul's Cross Churchyard*. Occasional Publications of the Bibliographical Society.

Ross, S. (2000) *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*. NPO/British Library, Occasional Publication. Available at: https://pdfs.semanticscholar.org/45ef/6351e887f4ee575c96bde5b5d1c55825dd4c.pdf

A new Roma (beta): a rich interface for ODD customization

Keywords: ODD, customization, user interfaces

The TEI module for writing or customizing an XML markup language (the “tagdocs” module of chapter 22, “Documentation Elements”), as well as a file defining or customizing an XML language using these elements, is typically referred to as One-Document-Does-it-all or ODD. The procedural nature of these elements makes it possible to develop form-based user interfaces to generate ODD; indeed, the current tool for customizing the TEI schema, Roma (http://roma.tei-c.org), is widely used by the TEI community to generate project-specific schemata. Roma, however, does not fully reflect the possibilities of the ODD subset of elements, and has been afflicted by a number of bugs and issues, many of which remain unresolved and are too complex to be fixed by the TEI Technical Council.

The TEI Technical Council has opted to create a replacement for Roma from scratch, using modern web technologies, expanding the number of features supported, and focusing on user interface. This new version of Roma (beta) is available at https://romabeta.tei-c.org). This effort is led by council member Raffaele Viglianti, who has developed a web application using the popular front-end libraries React and Redux. The full council provides regular feedback on both architectural decisions and the user interface. Just like the current version of Roma, this new version relies on the TEI Stylesheets to perform ODD transformations, including from ODD to XML schema formats. To perform these transformations, Roma (beta) interacts with OxGarage (http://www.tei-c.org/oxgarage/), an on-line service for TEI transformations.

This poster will introduce motivations for developing Roma, describe its architecture, and demo its features. It will also focus on future plans for development, and provide a way of collecting feature requests from attendees at the conference.

Authors

Raffaele Viglianti is a Research Programmer at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, where he works on a number of digital humanities projects and is the Technical Editor for the Shelley-Godwin Archive.

Raffaele’s research revolves around digital editions and textual scholarship, with a focus on editions of music scores.

How to encode the unsaid with the TEI

Where does text begin and, above all, where does it end? Do we reach the end of the text where the characters on the paper end? At what point does the unsaid become tangible in literature, how does this find expression in the medial side and materiality of text, and above all how does the TEI deal with it? More precisely: What possibilities are there to encode the unspoken? Is this at all purposeful and does it not contradict the basic principle of the TEI, according to which the TEI encoding is to be considered detached from the representation and is to be removed from the interpretation?

Many texts in world literature play with the phenomenon of vagueness. These texts leave the statement in the approximate, they suggest things and thus hand over the narrative to the reader's imagination. The narrative communication succeeds in these cases through the unity of language, text statement and media representation, which can differ depending on the printed edition. The reader of Theodor Fontane's *Effi Briest* "knows" what happens between Effi and Crampas in the dunes, even if it is not written anywhere and is only indicated by the length of the paragraph. In Arthur Schnitzler's *Reigen*, the dashes not only suggest, they speak for themselves. And the kiss exchanged in Paul Claudel's *L'Annonce faite à Marie*, which is immanently important for the plot, regularly falls through the grid, whereby the two kissers, who otherwise do not speak a word with each other, do not occur together in an automatically extracted communication graph.

The power and relevance of these blanks (“Leerstellen”) in literature is undisputed. It is only while reading that the recipient, drawing on his own knowledge of the world and experience, redeems the potential effect offered by the text and turns it into a literary work of art.[1] Only the cognitive challenge of the reader through the lack of unambiguity leads to the active execution of the action. The dialectic of pronunciation and concealment stimulates the assignment of meaning. How can TEI, as a tool for capturing sign sequences, text structure and abstract properties, deal with this phenomenon? Can it do justice to the unsaid and is that its task?

Based on the three works mentioned, the phenomenon of the unsaid in handwriting, printing and the respective TEI encoding is to be investigated. Theoretical considerations on the relationship between content, form and medium of text in comparison to non-text within the framework of the TEI are raised.

[1] Cf. Wolfgang Iser: Die Appellstruktur der Texte. Unbestimmheit als Wirkungsbedingung literarischer Prosa. Konstanz 1970. Wolfgang Iser: Der implizite Leser. Kommunikationsformen des Romans von Bunyan bis Beckett. München 1972.

Distant Spectators: Mining TEI-encoded periodicals of the Enlightenment

Alexandra Fuchs, Bernhard Geiger, Elisabeth Hobisch, Philipp Koncar, Sanja Saric, Martina Scholger

The poster will present the idea behind and first steps in the recently started project *Distant Spectators: Distant reading for periodicals of the Enlightenment*. The objective of this project is the application of distant reading and text-mining methods (topic modeling, meme diffusion, stylometry, sentiment analysis, network analysis) to the Spectators press, a journalistic genre of the 18th century Enlightenment, and the combination of these methods with the already existing expertise gained from close reading. This will provide an insight into the formation of trans-European ideas, literary techniques and cultural practices by employing quantitative methods to investigate authorship attribution, editorial networks, distribution of topics, transfer of micro-narratives etc.

The project builds on an existing and ongoing digital edition project, *The Spectators in the International Context*, which has been running since 2008 (https://gams.uni-graz.at/spectators). Currently it incorporates approximately 4000 individual texts in six languages (French, Italian, Spanish, English, German, Portuguese) with more than 9 million tokens. The discourses are encoded in TEI, representing the text structure and the narrative forms (e.g. reader's letter, fable, dreams) in the texts, and building registers of names, works, and places.

The TEI encoding builds the basis for the computational analysis. The benefit of the application of quantitative methods on the basis of an elaborate TEI model is the flexibility in building collections drawing on the affiliation of single texts to specific journals, to certain time periods, to individual keywords, etc. encoded in the TEI Header. Furthermore, specific textual structures and narrative features can be extracted and analyzed in relation to the entire corpus. A particular challenge is the compilation of a representative corpus for the application of quantitative methods due to the a) multilingual text-corpus, b) brevity of single discourses, and c) short period of publishing.

The main objective is to investigate how and which quantitative methods prove useful for the analysis of this multilingual corpus from the 18th century.

A new Roma (beta): a rich interface for ODD customization

Keywords: ODD, customization, user interfaces

The TEI module for writing or customizing an XML markup language (the “tagdocs” module of chapter 22, “Documentation Elements”), as well as a file defining or customizing an XML language using these elements, is typically referred to as One-Document-Does-it-all or ODD. The procedural nature of these elements makes it possible to develop form-based user interfaces to generate ODD; indeed, the current tool for customizing the TEI schema, Roma (http://roma.tei-c.org), is widely used by the TEI community to generate project-specific schemata. Roma, however, does not fully reflect the possibilities of the ODD subset of elements, and has been afflicted by a number of bugs and issues, many of which remain unresolved and are too complex to be fixed by the TEI Technical Council.

The TEI Technical Council has opted to create a replacement for Roma from scratch, using modern web technologies, expanding the number of features supported, and focusing on user interface. This new version of Roma (beta) is available at https://romabeta.tei-c.org). This effort is led by council member Raffaele Viglianti, who has developed a web application using the popular front-end libraries React and Redux. The full council provides regular feedback on both architectural decisions and the user interface. Just like the current version of Roma, this new version relies on the TEI Stylesheets to perform ODD transformations, including from ODD to XML schema formats. To perform these transformations, Roma (beta) interacts with OxGarage (http://www.tei-c.org/oxgarage/), an on-line service for TEI transformations.

This poster will introduce motivations for developing Roma, describe its architecture, and demo its features. It will also focus on future plans for development, and provide a way of collecting feature requests from attendees at the conference.

Authors

Raffaele Viglianti is a Research Programmer at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, where he works on a number of digital humanities projects and is the Technical Editor for the Shelley-Godwin Archive.

Raffaele’s research revolves around digital editions and textual scholarship, with a focus on editions of music scores.

How to encode the unsaid with the TEI

Where does text begin and, above all, where does it end? Do we reach the end of the text where the characters on the paper end? At what point does the unsaid become tangible in literature, how does this find expression in the medial side and materiality of text, and above all how does the TEI deal with it? More precisely: What possibilities are there to encode the unspoken? Is this at all purposeful and does it not contradict the basic principle of the TEI, according to which the TEI encoding is to be considered detached from the representation and is to be removed from the interpretation?

Many texts in world literature play with the phenomenon of vagueness. These texts leave the statement in the approximate, they suggest things and thus hand over the narrative to the reader's imagination. The narrative communication succeeds in these cases through the unity of language, text statement and media representation, which can differ depending on the printed edition. The reader of Theodor Fontane's *Effi Briest* "knows" what happens between Effi and Crampas in the dunes, even if it is not written anywhere and is only indicated by the length of the paragraph. In Arthur Schnitzler's *Reigen*, the dashes not only suggest, they speak for themselves. And the kiss exchanged in Paul Claudel's *L'Annonce faite à Marie*, which is immanently important for the plot, regularly falls through the grid, whereby the two kissers, who otherwise do not speak a word with each other, do not occur together in an automatically extracted communication graph.

The power and relevance of these blanks (“Leerstellen”) in literature is undisputed. It is only while reading that the recipient, drawing on his own knowledge of the world and experience, redeems the potential effect offered by the text and turns it into a literary work of art.[1] Only the cognitive challenge of the reader through the lack of unambiguity leads to the active execution of the action. The dialectic of pronunciation and concealment stimulates the assignment of meaning. How can TEI, as a tool for capturing sign sequences, text structure and abstract properties, deal with this phenomenon? Can it do justice to the unsaid and is that its task?

Based on the three works mentioned, the phenomenon of the unsaid in handwriting, printing and the respective TEI encoding is to be investigated. Theoretical considerations on the relationship between content, form and medium of text in comparison to non-text within the framework of the TEI are raised.

[1] Cf. Wolfgang Iser: Die Appellstruktur der Texte. Unbestimmheit als Wirkungsbedingung literarischer Prosa. Konstanz 1970. Wolfgang Iser: Der implizite Leser. Kommunikationsformen des Romans von Bunyan bis Beckett. München 1972.

Workshops

An introduction to Schematron and Schematron QuickFix

David Maus, Herzog August Bibliothek Wolfenbüttel

Keywords: Schematron, XML, Validation, Schematron QuickFix, Workshop

1 Schematron is a rule based validation language for structured documents. It was designed by Rick Jelliffe in 1999 (Jelliffe 1999) and standardized as ISO/IEC 19757-3 in 2006 (ISO 2006). The key concepts of Schematron validation are *patterns* that are the focus of a validation, *rules* selecting the portions of a document contributing to the pattern, and *assertion tests* that are run in the context of a rule. Schematron uses *XPath* both as the language to select the portion of a document and as the language of the assertion tests. This use of *XPath* gives Schematron the flexibility to validate arbitrary relationships and dependencies of information items in a document.

2 What also sets Schematron apart from other languages is that it encourages the use of natural language descriptions targeted to human readers. This way validation can be more than just a binary distinction (document valid/invalid) but also support authors of in-progress documents with quick feedback on erroneous or unwanted document structure and content.

3 The flexibility and (relative) simplicity of Schematron makes it an invaluable tool for XML-based text encoding projects. The range of supported tasks reaches from "hard" validation to enforce constraints on documents to "soft" validation to report potential problems such as Unicode characters from Unicode Private Use Areas to interactive error correction with Schematron extensions like Schematron QuickFix (Kutscherauer and Nadolu 2018)

4 This *full day workshop* will introduce the participants to principle idea of Schematron and discuss its application to XML-based text encoding projects. Together we will explore patterns, rules, and assertions as the basic Schematron concepts and touch phases, variables, and abstract patterns

as more advanced features of Schematron validation. The workshop will end with Schematron QuickFix, an extension to Schematron that allows for interactive error corrections in XML editing environments like <oXygen/>.

5 From the participants the workshop requires a general understanding of XML document editing and basic knowledge of *XPath*. The material requirements are a projector and laptops to follow through with the examples given in the workshop. For the main part – basic and advanced Schematron concepts – any operating system with a recent Java Runtime is sufficient, the introduction to Schematron QuickFix will use the <oXygen/> XML editor.

6 Participants are recommended to bring their own device.

BIBLIOGRAPHY

International Organization for Standardization. 2006. Part 3: Rule-Based Validation - Schematron. ISO/IEC 19757-3:2006(E).

Jelliffe, Rick. 1999. Using XSL as a Validation Language. https://web.archive.org/web/20000415135808/http://www.ascc.net:80/xml/en/utf-8/XSLvalidation.html.

Kutscherauer, Nico, and Octavian Nadolu. 2018. Schematron Quick Fixes Specification. Draft March 2018. Quick-fix support for XML Community Group. https://schematron-quickfix.github.io/sqf.

AUTHOR

DAVID MAUS

Herzog August Bibliothek Wolfenbüttel

Minimalist TEI Publishing with CETEIcean (/sɪˈti:ʃn/)

Keywords: publishing, processing, static sites, minimal computing

This half-day workshop will introduce CETEIcean (pronounced /sɪˈti:ʃn/, similarly to the word “cetacean”), a JavaScript library for displaying TEI in a web browser. CETEIcean’s key feature is the isomorphic transformation of TEI as HTML by registering modified TEI elements with the browser using Custom Elements.[1] This is in contrast to typical transformations from TEI to HTML that stick with the semantically poor element set of HTML. Loading a TEI document with CETEIcean will make it fully usable by the browser: TEI elements can be styled with CSS and manipulated for interaction with JavaScript just like HTML, thus avoiding XSLT transformation steps (see Cayless and Viglianti 2018). CETEIcean’s approach ows a lot to TEI Boilerplate (http://teiboilerplate.org/), but it is based on web standards and is more flexible. It does not require an in-browser XSLT step, nor any modification to the source XML; TEI content can be loaded in the browser via a JavaScript fetch operation, or via server-side processing.

This workshop will also introduce “minimal computing” principles, particularly as pioneered by the GO:DH Minimal Computing Working Group.[2] This approach fosters a “reconnecting with our knowledge production” in order to think critically about the question “what do we need?” (Gil 2015). Identifying the minimal technical stack for running a TEI project can help content creators to both reach their goals quickly and effectively, but more importantly they might be able to do so even when there are unwanted constraints in the resources they can access. Minimal computing principles argue that this can be an essential instrument in giving voice to underfunded and underrepresented groups.

Many scholarly publications powered by the TEI rely on server side infrastructure for publication, typically by leveraging XML technologies such as XSLT and XQuery (via XML databases). Adopting a minimal computing stance and CETEIcean may reduce if not remove reliance on these technologies; yet the capability, flexibility, and established practice of these tools is undeniable, so why use a different approach?

- **Web technologies for web publication.** XSLT/XQuery are not sufficient, by themselves, to create a digital publication on the web, they are simply a means to transforming TEI into HTML. But to create a fully featured publication on the web, HTML needs to be supported by CSS for styling and JavaScript for user interaction. What if we

[1] https://html.spec.whatwg.org/multipage/custom-elements.html#custom-elements.
[2] http://go-dh.github.io/mincomp.

could skip the transformation part and focus on the web technologies needed for a digital publication?

- **Getting started faster.** By focusing on web technologies, new TEI users will not have to learn XSLT as well as the TEI to publish their first document.
- **Semantics.** Converting TEI to HTML is the most common and most practical way of publishing TEI texts on the web, but HTML lacks what TEI has: a very well-considered and mature set of semantic tags for encoding texts. When converting TEI to HTML the semantic distinctions in the markup are often lost in favor of typographic distinctions in the display. In other words, the data model represented in the TEI fails to carry over to the online version.
- **Preservation.** XSLT/XQuery transformations are often performed "on the fly" by server-side technology to provide data to a front-end application (written in HTML, CSS, and JavaScript). By using CETEIcean, TEI can be built into "static sites" that do not require server-side computation, making it easier to preserve the application into the future.

Syllabus

This half-day workshop will cover the following topics.

- Introduction to Minimal Computing principles and discussion.
- Introduction to CETEIcean and motivations.
- Using CETEIcean to build a static site from a provided HTML template. We will discuss the structure of the template and how to change it for future use.
- Using CSS to style CETEIcean TEI. No previous knowledge of CSS required.
- Adding user interactivity to your TEI via CSS and simple JavaScript. We will discuss a number of examples and address questions/requests from the attendees.
- Publishing online. We will look at simple ways of publishing a CETEIcean-powered site online for free via GitHub.

Tutors

Raffaele Viglianti is a Research Programmer at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, where he works on a number of digital humanities projects and is the Technical Editor for the Shelley-Godwin Archive. Raffaele's research revolves around digital editions and textual scholarship, with a focus on editions of music scores. Raffaele currently serves on the Technical Council of the Text Encoding Initiative.

Hugh Cayless has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

Bibliography

Cayless, Hugh, and Raffaele Viglianti (2018). “CETEIcean: TEI in the Browser.” Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In *Proceedings of Balisage: The Markup Conference 2018*. Balisage Series on Markup Technologies, vol. 21 (2018). https://doi.org/10.4242/BalisageVol21.Cayless01.

Gil, Alex: “The User, the Learner and the Machines We Make” (2015). Blog post at: http://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner/

1.

The *Wien[n]erisches Diarium Digital – Digitarium* is an ongoing project that aims at making available one of the oldest newspapers in the world available as a high quality full text. The *Wiennerisches Diarium*, now *Wiener Zeitung*, first appeared on 8th August 1703, initially with 2 issues per week of usually 8–12 pages each but reaching over 40 pages regularly by the second half of the 18th century. From October 1813, there were daily issues (including Sundays).

The project makes use of the Transkribus software and *Handwritten Text Recognition (HTR)* models trained specifically on the newspaper's issues to achieve a reasonably high quality full text – on average, less than 1.5% *character error rate (CER)* – from automated processing. During the first 2 years, the project implemented this automated workflow and improved the HTR models by making available 420 issues from 1703 – 1799 (5 per year where images were already available). In the end, the team wants to include all issues from 1703 until the 1940es in an extensive corpus with a versatile frontend that can cater to different research needs.

While a recent grant application (the outline of which had been presented in Tokyo) has not been successful, the project team still is intent on developing an interface together with researchers and the interested public alike. Several questions that arise from the serial nature of the source, the amounts of text involved as well as the linguistic changes over more than 200 years have to be addressed and combined with a user centred design approach so that the texts can be presented, read, searched and otherwise reused easily.

This one-day workshop wants to include the TEI community in this development. The first part will introduce participants to the project, its workflow and the current web interface. A short survey is to collect the initial reactions to the interface.

The second part will focus on research questions that can be answered by periodical texts and how both the framework in which they are presented and the encoding of the texts and their metadata can support a wide variety of research disciplines. Participants will be asked to try to answer a research question from a field of their choosing and record what steps they take, what functionality they would like to see included in the web frontend to help them in their research and whether, and if so, how, they would like to contribute in improving the quality of the source material.

Both parts will be connected by the presentation of (and comparison to) the results of a two day conference and several "annotate-a-thons" held from 2017 – 2019.

The results of the workshop will be discussed in the form of an article for the *JTEI* while also being an important basis for the further development of the framework used to present the *Diarium*'s texts (which of course is available as an open source: wdbplus). The texts will be made available separately later this year.

2.

Participants should have a laptop with a working internet connection so they can use the projects website. The room needs a means of projection. Those who want to use their own data and/or install the presentation tool themselves need to have eXist installed and should, if possible, already have wdbplus running.

While the *Diarium* texts are in German, knowledge of German is not necessary. However, if participants want to suggest a periodical in a different language for use in the project, this is highly welcome and will be included for use during the workshop.

Hands-on TEI publishing - workshop

Crossing the divide between encoded XML sources and tangible, published digital edition has always been a weak spot for TEI community. Efforts of the TEI Simple project aimed to bridge that gap with TEI Processing Model idea. TEI Publisher, an eXist-db based application brought the promises of TEI Simple to life with its implementation of the processing model enhanced with an app generator, allowing to create standalone digital editions out of the box. Publishing an edition from TEI sources would usually involve tedious work on complex stylesheets and significant effort in building an application on top of it. Using the TEI Processing Model, customising the transformation of the text is all done in TEI ODD which is fairly easy to grasp even for non-technical users. The power of the eXist-db database and the application framework on the other hand takes care of all the other core features like browsing, search and navigation. Since version 4 TEI Publisher provides a new abstraction layer for the application. Switching to a web components based design allowed us to introduce easy to customize HTML templates into the picture. This way adjusting the layout of the published edition can be reduced to moving around small blocks of HTML, not unlike playing with LEGOs.

Customising the appearance of the text in TEI ODD was proven to easily save thousands of lines of code for media specific stylesheets. New web component based templates offer similar advantages for the application layer. Thanks to a growing library of components to just plug into new applications - from faceted search to IIIF viewer - custom functionality can be easily added manipulating just HTML templates.

The proposed workshop intends to introduce the main concepts behind the new TEI Publisher and provide a tutorial on how to generate custom standalone edition using it. As an inspiration it will also present examples of real apps built with the TEI Publisher, including the challenge of recreating the recently released Van Gogh Letters edition within Publisher.

We hope that exposure to the concepts and technologies presented during the workshop will give its participants a point of exit in the task of publishing their own research data.

Tutors

Wolfgang Meier (wolfgang@existsolutions.com) open source developer, founder of eXist-db XML native database project, author of the TEI Publisher

Magdalena Turska (magdalena@existsolutions.com) developer, one of the authors of the TEI Simple and TEI Processing Model, TEI Technical Council and TAPAS advisory board member, working on integration of TEI Processing Model into TEI and TAPAS infrastructure

Audience

Expected audience is ca 20 participants, ranging from early stage scholars interested in self-publishing of their individual research, through archivists and curators of digital resources, to developers building applications for digital publication of scholarly content. Previous workshops attracted between 15 and 30 people, belonging predominantly to the first category, but with significant presence of two other groups as well. Please note we had TEI Publisher related workshops during previous TEI conferences but we believe that substantial changes in v4 together with few years break since last European workshop (Vienna 2016) merit a slot to talk about it again. Also the actual content of the workshop will substantially differ from previous editions.

Requirements

Spacious room with possibility to set up with participants' own laptops, plenty of power sockets and good Wi-Fi coverage, beamer and whiteboard. Possibility to rearrange tables/chairs for work in small groups would be appreciated. We'll want to be able to connect to eXist solutions server, so we'd need to check in advance if local firewall doesn't block such connections.

Format

Full-day workshop: ca 6h tuition plus lunch and coffee breaks

Authors:

Magdalena Turska, Wolfgang Meier

Demonstrations

MediaWiki TEI extension demo

Thomas Pellissier Tanon

MediaWiki is one of the leading open source wiki engines. It is well known for being used by Wikipedia but it also powers many other websites including Wikisource, the Wikipedia sister project dedicated to transcription, and the TEI wiki. MediaWiki default encoding for wiki pages content is wikitext. However the platform provides a powerful content abstraction system, allowing, e.g. to use an other content encoding for some or all wiki pages.

The aim of this demo is to present a new MediaWiki extension, called "TEI" that allows to create MediaWiki pages encoded in TEI P5 instead of Wikitext. This extension is a work in progress. It provides an implementation of a subset of TEI content tags (i.e. excluding the TEI header) inspired by TEI simplePrint. A simple XML editor is provided with validation and auto-complete based on a configurable ODD bundled with the extension. A beginning of WYSIWYG editor is also available. A demo wiki is updated daily with the latest version of the extension and provides some examples. Before the demonstration we plan to improve both the WYSIWYG and XML editor, implement the support of more TEI tags and attributes and improve the extension configuration and extensibility. If time permits, we also plan to integrate it with ProofreadPage, an other MediaWiki extension providing a transcription workflow for MediaWiki and currently based on Wikitext.

TEI as a Graph

Andreas Kuczera, Academy of Science and Literature, Mainz

As TEI is not a format, though many people think it is. It's a de facto standard that specifies Guidelines for document interchange. Actually the Guidelines are based on the XML but this is only one possible technical way of expressing the phenomenons. In the graph you can use multi-hierarchical annotations layers. Graph models are very easy to read and understand. So DH-People and "normal" scientists have a level of discussion in common. A Graph can be expressed as RDF so the step from a Graph to linked open data is easy to make.

In this paper a small xml-example in DTA-Base-Format will be imported into the graph-database neo4j and then be converted to the Standoff-Property-Json-Format[1] but this toolchain works for every TEI-XML-file.[2] The exported Standoff-Property-Json data can then be imported into the Standoff-Property-Editor SPEEDy, which can manage multi-hierarchical annotations. Standoff-Formats are well-known but they have some limitations. So you are not allowed to change the base text (datum) after having started with the annotations as the indexes would be damaged. In our system annotated documents can be edited as the indexes are recalculated when the document is saved.

Convert DTA-XML with neo4j to Standoff Property JSON

In a first step we import a small xml-example into a neo4j (https://neo4j.com) instance using apoc.import.xml (https://github.com /neo4j-contrib/neo4j-apoc-procedures-function)

The example is one folio (https://seafile.rlp.net/f/6282a26504cc4f079ab9/?dl=1) from the DTA (https://www.deutschestextarchiv.de). Here you can find the XML-Testfile and this is the Link (http://www.deutschestextarchiv.de/patzig_msgermfol841842_1828/11) to the DTA-Version.

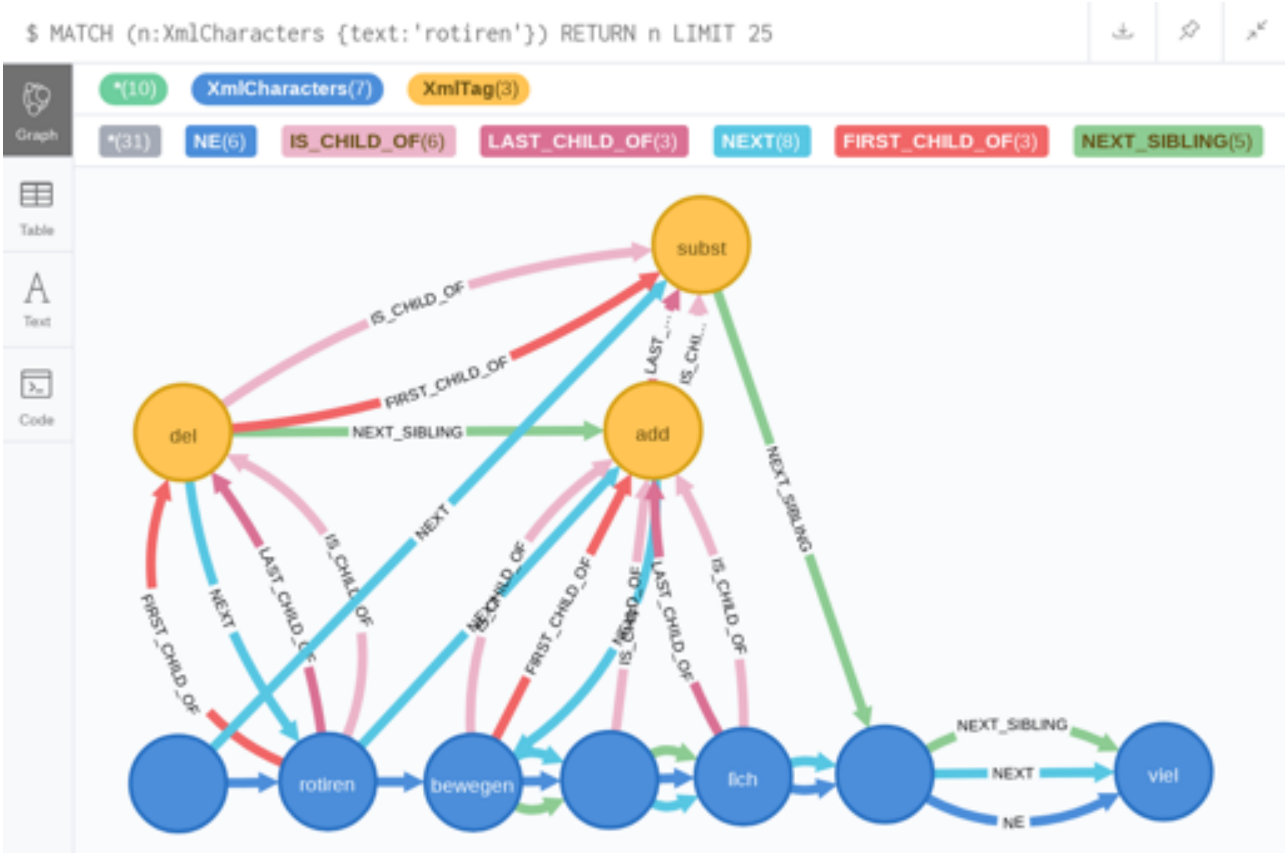
Import into neo4j

The import into neo4j runs with:

```
// Import xml-example from DTA to neo4j
call
apoc.xml.import('https://seafile.rlp.net/f/6282a26504cc4f079ab9/?dl=1',
{connectCharacters: true, charactersForTag:{lb:' '},
filterLeadingWhitespace: true}) yield node
return node;
```

[1] The Standoff Property Format is explained in detail in Iian Neill, Andreas Kuczera: The Codex – an Atlas of Relations. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) text/html Format. DOI: 10.17175/sb004_008.

[2] The example is the xml export of folio 11 of the notes of Gotthilf Friedrich Patzig about Humboldts Kosmos-Lecture accessible in the German Textarchive (http://www.deutschestextarchiv.de/book/show/30962).

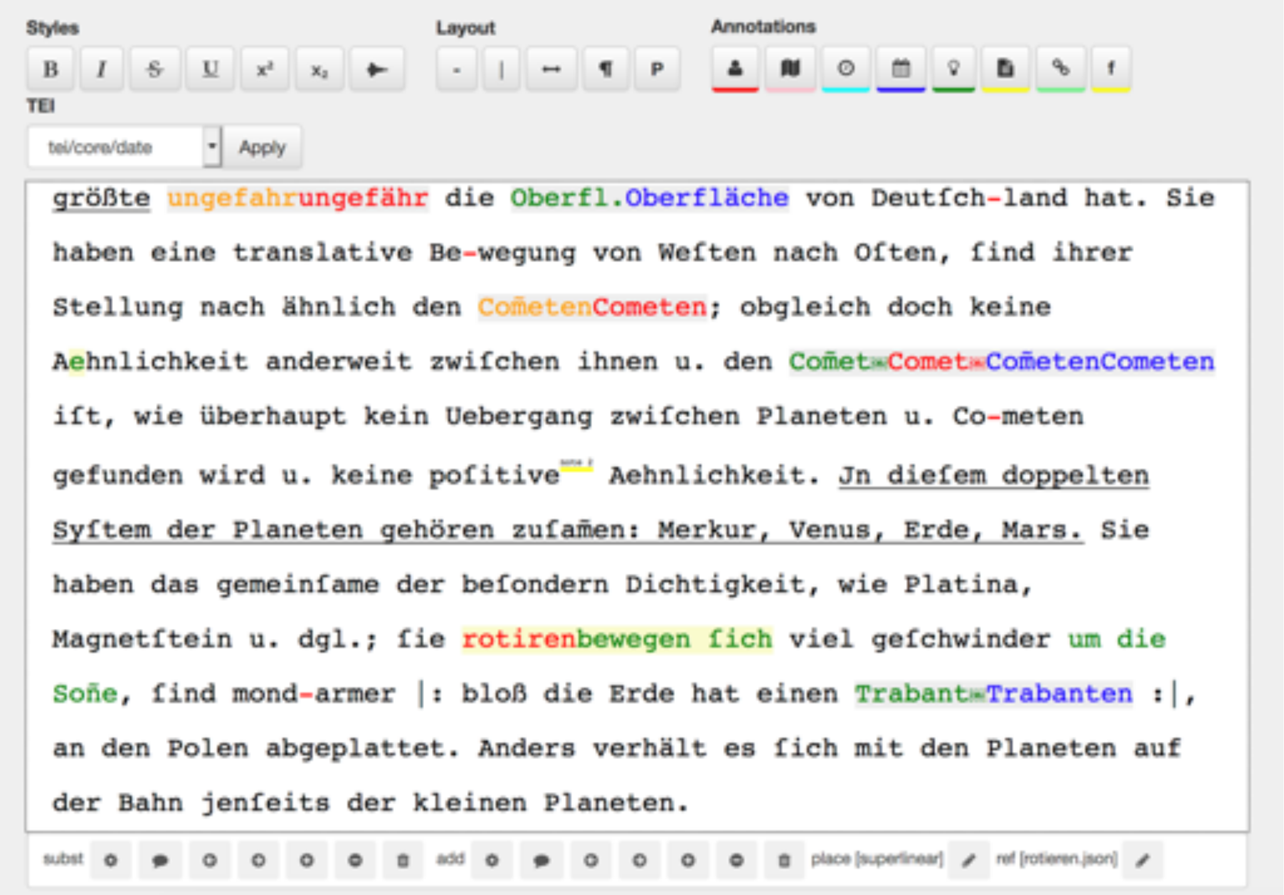


```
$ MATCH (n:XmlCharacters {text:'rotiren'}) RETURN n LIMIT 25
```

Figure1: TEI-XML-Example in neo4j (Kuczera).

In Figure 1 you can see a snippet of the example in the Graph-Database. In this import to the graph-database the xml-file is imported as an xml-tree with the root-element at the top level. The hierarchy of the xml is expressed with IS_CHILD_OF, FIRST_CHILD_OF, LAST_CHILD_OF etc. edges connecting all elements which are converted to nodes of type XmlTag for the elements or XmlCharacter for the text. The seriality of the XML-file is expressed by NEXT, which make reexporting XML possible. In addition all text nodes are connected by NE edges, connecting all text without any elements in between. Whitespaces become a textnode on their own. The example shows that importing a DTA-Baseformat-XML-File keeps all informations from the xml-version and re-exporting to xml is possible.

Export from neo4j to Standoff Property JSON



größte ungefahrungefähr die Oberfl.Oberfläche von Deutſch-land hat. Sie haben eine translative Be-wegung von Weſten nach Oſten, ſind ihrer Stellung nach ähnlich den Com̄etenCometen; obgleich doch keine Aehnlichkeit anderweit zwiſchen ihnen u. den Com̄etCometCom̄etenCometen iſt, wie überhaupt kein Uebergang zwiſchen Planeten u. Co-meten gefunden wird u. keine poſitive Aehnlichkeit. Jn dieſem doppelten Syſtem der Planeten gehören zuſam̄en: Merkur, Venus, Erde, Mars. Sie haben das gemeinſame der beſondern Dichtigkeit, wie Platina, Magnetſtein u. dgl.; ſie rotirenbewegen ſich viel geſchwinder um die Soñe, ſind mond-armer |: bloß die Erde hat einen TrabantTrabanten :|, an den Polen abgeplattet. Anders verhält es ſich mit den Planeten auf der Bahn jenſeits der kleinen Planeten.

Figure2: TEI-as-a-Graph in the Standoff-Property-Editor SPEEDy (Kuczera).

The next step is to export the data with some cypher to the Standoff-Property JSON-Format, which can be directly copied out of the neo4j-browser-window. This json can then be imported in the [SPEEDy (https://github.com/argimenes/standoff-properties-editor)] Standoff Property Editor which can be found on GitHub (https://github.com/argimenes/standoff-properties-editor).

In the README-Section of the SPEEDy Github Repo you can find a Link (https://argimenes.github.io/standoff-properties-editor/) to the Test-Instance hosted on Github-Pages. We have prepared the example in SPEEDy. Just select „TEI-XML → SPEEDY IV" in the file-Section and load the data. Below the Editor-Window you can press the UNBIND-Button and inspect the exported json in the window below.

Figure 2 shows the results of the conversion without any further treatment by hand. The plain text is the result of the xml-file with all elements deleted and not very good to read. But if you select a part of the text the according annotations are shown below the editor window, so the semantic is not lost.

Further steps will be some algorithms to put deleted text in an annotation of the added text to get a readable text which then can be annotated further. Another task is developing an export function to xml. Another approach could be to do the refactoring of the xml in the graph-database to get clean Standoff-Data out of the Graph-DB. From my point of view TEI as Graph can be the next technical step for TEI to get better support and linking to Linked Open Data projects and to overcome the uni-dimensional restriction of xml.
I want to say thanks to Stefan Armbruster from neo4j for the export-cypher-query and the

implementation of the XML-Import funkctions to apoc (https://github.com/neo4j-contrib/neo4j-apoc-procedures-function) and Iian Neill for his work on SPEEDy (https://argimenes.github.io/standoff-properties-editor/).

TEITOK – TEI based annotated corpora

Maarten Janssen

Biography

Maarten Janssen is a researcher at ÚFAL – Charles University, Prague, and the author of the TEITOK corpus platform, a TEI based tool for creating, maintaining, and distributing annotated corpora. TEITOK is used in a growing number of corpora around the world, primarily for historical, spoken, and learner corpora. He is directly involved in a number of TEITOK based corpus projects, including COPLE2, PEAPL, PostScriptum, CORDEREGRA, EFFE-ON, CzeSL, and CoDiaJE.

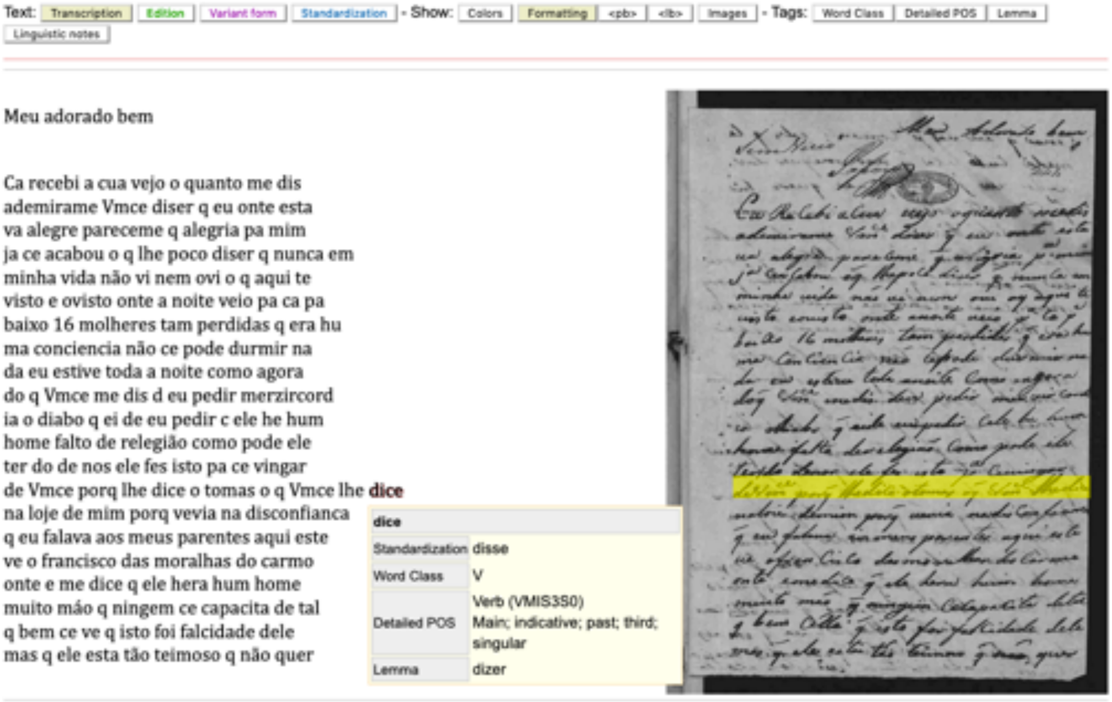
Keywords: Annotated corpora, Dependency grammar, POS tagging, TEI visualization

Abstract

TEITOK is a web based tool building, annotating, and distributing corpora, in which corpus files are stored in TEI/XML. It combines the needs of those how want to do detailed philological markup with the requirements of a searchable, annotated linguistic corpus, and is being used in a growing number of corpora around the world, primarily for historical, spoken, and learner corpora.

Proposal

TEITOK is a web based tool building, annotating, and distributing corpora, in which corpus files are stored in TEI/XML. It combines the needs of those how want to do detailed philological markup with the requirements of a searchable, annotated linguistic corpus, and is being used in a growing number of corpora around the world, primarily for historical, spoken, and learner corpora.

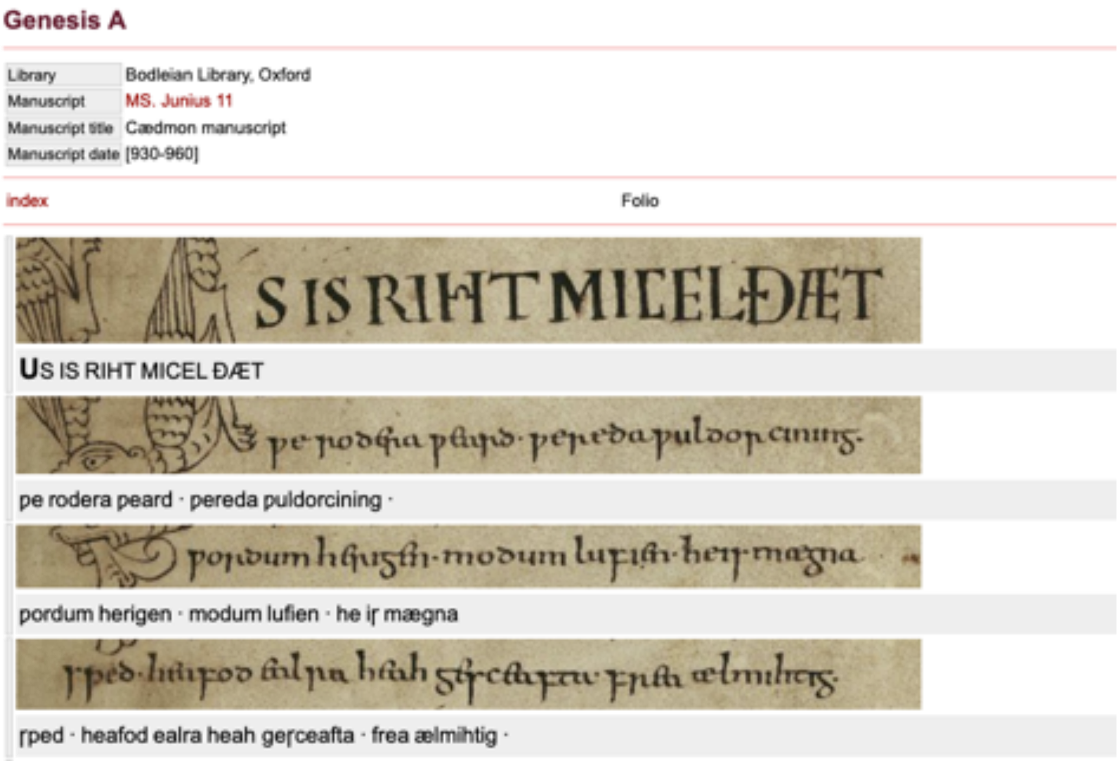


Meu adorado bem

Ca recebi a cua vejo o quanto me dis
ademirame Vmce diser q eu onte esta
va alegre pareceme q alegria pa mim
ja ce acabou o q lhe poco diser q nunca em
minha vida não vi nem ovi o q aqui te
visto e ovisto onte a noite veio pa ca pa
baixo 16 molheres tam perdidas q era hu
ma conciencia não ce pode durmir na
da eu estive toda a noite como agora
do q Vmce me dis d eu pedir merzircord
ia o diabo q ei de eu pedir c ele he hum
home falto de relegião como pode ele
ter do de nos ele fes isto pa ce vingar
de Vmce porq lhe dice o tomas o q Vmce lhe dice
na loje de mim porq vevia na disconfianca
q eu falava aos meus parentes aqui este
ve o francisco das moralhas do carmo
onte e me dice q ele hera hum home
muito máo q ningem ce capacita de tal
q bem ce ve q isto foi falcidade dele
mas q ele esta tão teimoso q não quer

dice	
Standardization	disse
Word Class	V
Detailed POS	Verb (VMIS3S0) Main; indicative; past; third; singular
Lemma	dizer

With regards to textual mark-up, it allows the visualisation of TEI documents directly in a browser, using CSS and JavaScript to visualize the different TEI elements in a customisable way. It can display facsimile images alongside the text, and has additional display options for specific types of TEI documents, such as a line-by-line visualisation for aligned facsimile transcriptions, and a view including a waveform display for time-aligned audio transcriptions.



Genesis A

Library	Bodleian Library, Oxford
Manuscript	MS. Junius 11
Manuscript title	Cædmon manuscript
Manuscript date	[930-960]

index Folio

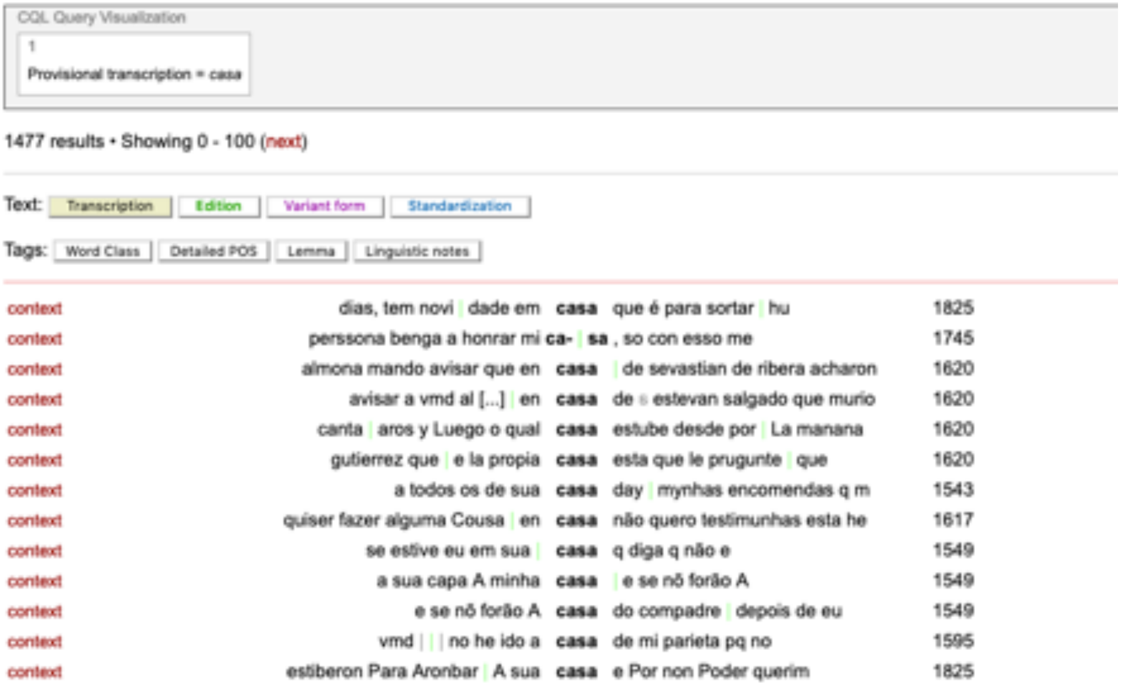
Us IS RIHT MICEL ÐÆT

pe rodera peard · pereda puldorcining ·

pordum herigen · modum lufien · he iꞅ mægna

ꞅped · heafod ealra heah geꞅceafta · frea ælmihtig ·

With regards to linguistic annotation, it allows TEI documents to be tokenised inline, after which each token can be adorned with information such as POS, lemma, or dependency relations. And the tokenised corpus can then be automatically exported as a linguistic corpus using the Corpus Workbench corpus tool, making it possible to search through the corpus using its expressive search languages. Different from most corpus search interfaces, TEITOK displays TEI/XML fragments in the search results, including hence the full textual mark-up of the source document.



CQL Query Visualization
1
Provisional transcription = casa

1477 results • Showing 0 - 100 (next)

Text: Transcription Edition Variant form Standardization
Tags: Word Class Detailed POS Lemma Linguistic notes

context	dias, tem novi | dade em	**casa**	que é para sortar | hu	1825
context	perssona benga a honrar mi **ca-** |	**sa**	, so con esso me	1745
context	almona mando avisar que en	**casa**	| de sevastian de ribera acharon	1620
context	avisar a vmd al [...] | en	**casa**	de s estevan salgado que murio	1620
context	canta | aros y Luego o qual	**casa**	estube desde por | La manana	1620
context	gutierrez que | e la propia	**casa**	esta que le prugunte | que	1620
context	a todos os de sua	**casa**	day | mynhas encomendas q m	1543
context	quiser fazer alguma Cousa | en	**casa**	não quero testimunhas esta he	1617
context	se estive eu em sua |	**casa**	q diga q não e	1549
context	a sua capa A minha	**casa**	| e se nõ forão A	1549
context	e se nõ forão A	**casa**	do compadre | depois de eu	1549
context	vmd | | | no he ido a	**casa**	de mi parieta pq no	1595
context	estiberon Para Aronbar | A sua	**casa**	e Por non Poder querim	1825

For tokenised corpora, it also allows storing multiple orthographic realisations for each token, such as a semi-palaeographic transcription and a regularised orthography, which can then in turn be used in the document view to display various editions of the same document. The textual metadata can be used in a number of different ways, for instance to display all the documents in the corpus on the world map. And the combination of metadata and token-based annotation allows for detailed corpus research on richly annotated documents.

Character Counting

This is demonstration of a character counting system. Character counting is not particularly new, is not all that interesting, and is not particularly difficult. And even the added fact that the output is a useful table of the characters found in an input file, sortable by frequency, character, Unicode code point, or (perhaps uselessly) by Unicode name is not particularly remarkable.

But add to that feature set the fact that the system works by running an XSLT program that writes an XSLT program, and it starts to get interesting. Furthermore, although the input file can be any XML document, the system will semi-intelligently handle several different kinds of input, currently including TEI, WWP, XHTML, and yaps. (That list may change before the conference — e.g., I am likely to add DocBook or JATS.)

In all cases attribute values can be included or excluded and whitespace can be normalized, ignored, or left as is at user option via a parameter. If the system knows the input language, further parameters may be specified to control whether or not metadata is included and perhaps other details (like choosing <corr> over <sic>). Lastly, the system performs a lookup into the Unicode database to get the correct Unicode name of each character.

Author: Syd Bauman

Recogito: from Semantic Annotation to Digital Scholarly Edition

Rainer Simon

Gimena del Rio Riande (IIBICRIT, CONICET), (AIT), Elton Barker (The Open University), Leif Isaksen (University of Exeter), Rebecca Kahn (*Alexander von Humboldt Institut für Internet und Gesellschaft* HIIG), Valeria Vitale (University of London), Antonio Rojas Castro (Cologne Center for eHumanities, Universität zu Köln), Hugh Cayless (Duke University)

Recogito (https://recogito.pelagios.org/) is a web-based environment for collaborative semantic annotation, developed by Pelagios (https://commons.pelagios.org). It is open source software, supports plaintext (.txt extension) as well as TEI/XML encoded text (.xml extension), and allows users to export the results of their work in different formats, including RDF, TEI/XML, GeoJSON, etc. Originally, the tool has been designed with a focus on scholarly geographic annotation, i.e. the transcription, marking up and geo-resolving of geographical documents such as itineraries, maps and travel reports. More recently, however, the feature set was expanded in order to provide more general annotation functionality. Perhaps the most notable feature of Recogito is the ability to produce semantic markup without the need to work with formal languages directly. Through an easy-to-use interface, users can navigate digitized documents; create personal collections; add tags and comments; build up tagging vocabularies, and geo-resolve place references by linking them to gazetteers. Users can either work alone in a closed workspace, or together as groups of collaborators. Recogito also makes it easy to apply Named Entity Recognition (NER) to TEI documents, with the possibility to choose between different recognition engines and authority files for entity resolution.

In this demo we will show how Recogito can serve as a useful environment for the efficient creation of minimal digital editions. Starting from plaintext source files, we will demonstrate the workflow for uploading content, creating semantic annotations, exporting to TEI, refining the markup, and publishing the results as an online digital edition. As a case study, we will present a geographically annotated corpus of early Argentinian texts. This edition was produced by semantically enriching sources with references to an early colonial american gazetteer, funded in part through a Pelagios Resource Development Grant in 2017.

Demonstration: TEI Graz 2019

Creating and implementing an ontology of documents and texts in the Textual Communities project.

First, I thank the conference organizers for allowing me to present the principles and practice of the Textual Communities project. Textual Communiies is now fully deployed and functioning. Time does not permit me, at the conference, both to describe the principles of Textual Communities and to show what Textual Communities does what it does. Accordingly, in this document I present the principles behind the project – essentially, built on a novel and comprehensive theory of What Text Is. In the demonstration, I will briefly refer to these principles and spend the rest of the talk showing Textual Communities.

In this principles paper, I want to focus now on two aspects of what we have done. The first is intellectual, the second social. First, the intellectual. In 2008, when I was thinking about what a cooperative and dynamic scholarly editing system might need, I went to a meeting at the University of Edinburgh on the marriage of Mercury and Philology. I outlined my ideas then of how such a system might be made, and a computer scientist there (whose name I don't recall) said to me: how you make it is not important. What is important is that you have the right model. Get that right, and everything follows. That set me to thinking. In the business of scholarly editing, there are three words which matter above all others: Document. Work. Text. What, exactly, do these three mean? How, exactly, do they relate to each other?

Like all who have lived under the shadow of the TEI, I have carved in me the formula of the Durand, Mylonas, Renear and DeRose "what is text really" article, which famously declares that text – any text – is composed of an ordered hierarchy of content objects. A tree, in fact. Dante's *Commedia* is made up of three canticles, each divided into thirty-three or thirty-four cantos, each canto divided into around 140 lines, give or take. Right back in the earliest days of the TEI, Michael and Claus showed how the Wittgenstein Nachlass could be represented according to this model, and I followed their example in my own work on medieval English texts. This model, I now see, is exactly half right.

Over the last decade, I've been engaged in an intermittent and occasionally fervent discussion with Peter Shillingsburg, Paul Eggert, Hans-Walter Gabler and Barbara Bordalejo about the meaning of these terms: document, text, work. We are editors. We deal with texts – but also, we deal with documents. In fact, every text we ever touch exists in a document. We have no way to the text except through a document.

Of course, the authors of What is text really, and the creators of the TEI, knew that documents and texts are intricately related. They knew that the *Commedia* exists in manuscripts and printed books. Accordingly, the TEI created a system of representing the physical characteristics of documents by the use of milestone elements: indicators of page, column and line breaks interspersed through text. Many of us found this quite adequate, and large quantities of text were encoded with this model.

We always knew there were problems with this model. Usually, this problem is expressed as “overlapping hierarchies”. The text itself is encoded in nicely ordered divisions: but the breaks between those divisions usually do not correspond with the page and other breaks. The text is one hierarchy, the document elements are scattered through that hierarchy. We can see that in this simple example:

```
<text>
        <body>
        <pb n="1r"/>
                <div n="Sample" type="entity">
                        <lb/><ab n="1">This line
                        <lb/>runs across several
                        <lb/>lines</ab><ab n="2"> While this
                        <lb/>block runs
        <pb n="1v">
                        <lb/>across a
                        <lb/>page break.
                        </ab>
                </div>
        </body>
</text>
```

Figure 1. A simple text with an “overlapping hierarchy”

So far, I have been talking about text as if it is – well, something we find in documents, or in what we say to each other. Usually, that is what we mean: we collapse all three terms work, text, and document so that the one word, text, might stand for all three. Over time, I have come to realize that this is not just a simplification, it is profoundly wrong.

Paul Eggert is a most hospitable person, and is fond of inviting his fellow textual scholars to come visit him in Canberra. Should you accept, he will take you for a walk in the Mount Ainsley park above his house. He will stop you in front of a tree, like this one, point to the marks, and ask you: is this a text?

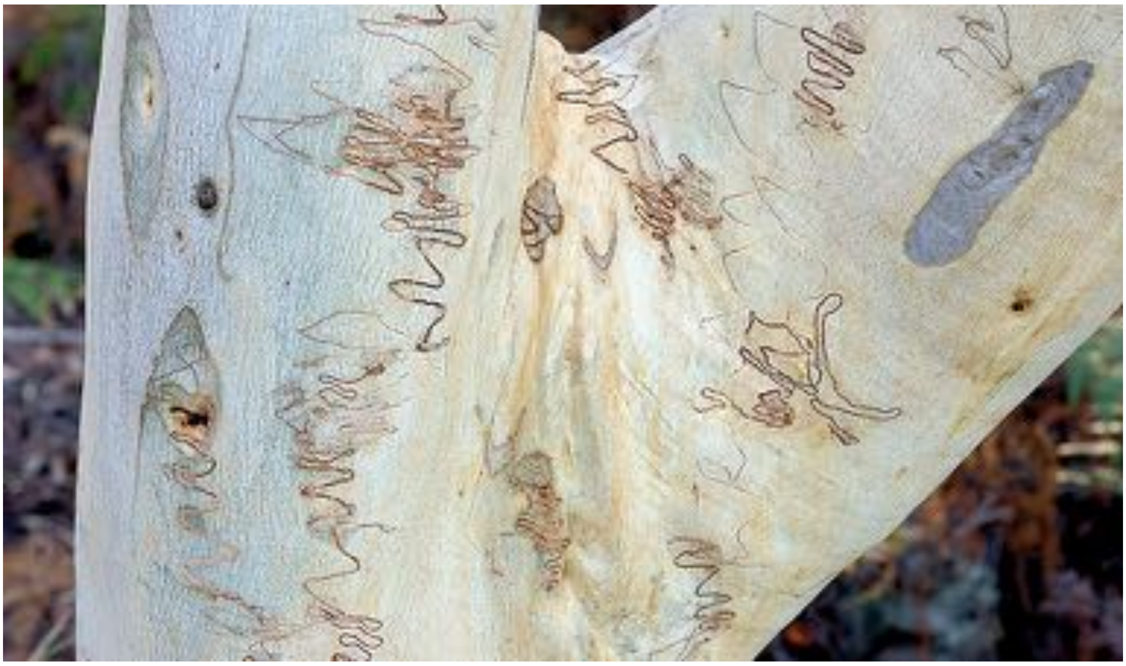


Figure 2. A “scribbly gum” tree.

Well, you think: these marks could be a text. There could be a pattern. Perhaps here is a crude pictograph of an animal. It might be some relative of Thai. In fact, Paul will tell you, these are marks made by a grub burrowing under the bark. Later, the bark falls away and reveals these squiggles.

As Paul Eggert puts it: this is not a text because it does not represent a human communicative act. And here is the first part of our definition of a text: it must represent an act of human communication. One may be more precise yet: we are speaking of linguistic communication. If it does not represent an act of communication, it is not a text. It is just marks on paper, or scratchings on a tree.

One may go further, and assert that indeed, each act of communication may be represented exactly as Durand and his co-authors suggested: as an ordered hierarchy of content objects. Here is where they were half right. Sometimes the act of communication may have a very simple structure: a single word or sentence. Or it may be as complex as the Bible, or the Commedia. But for every act of communication there is, we may detect a structure.

In our system, we call these meaningful units of communication “entities”. We do not call them “works”: works is a loaded term in textual scholarship. So, a text must be an act of communication. But that is not all a text is. An act of communication must exist in a physical form: or it does not exist in all. It must be inscribed into a manuscript, or printed on pages in a book, or recorded on a tape. Even in our thoughts it has an actual physical presence as synapses in our brain open and close. The reverse is true. Marks in a document which do not represent an act of human communication are not a text. They are just marks. Hence, this definition of a text:

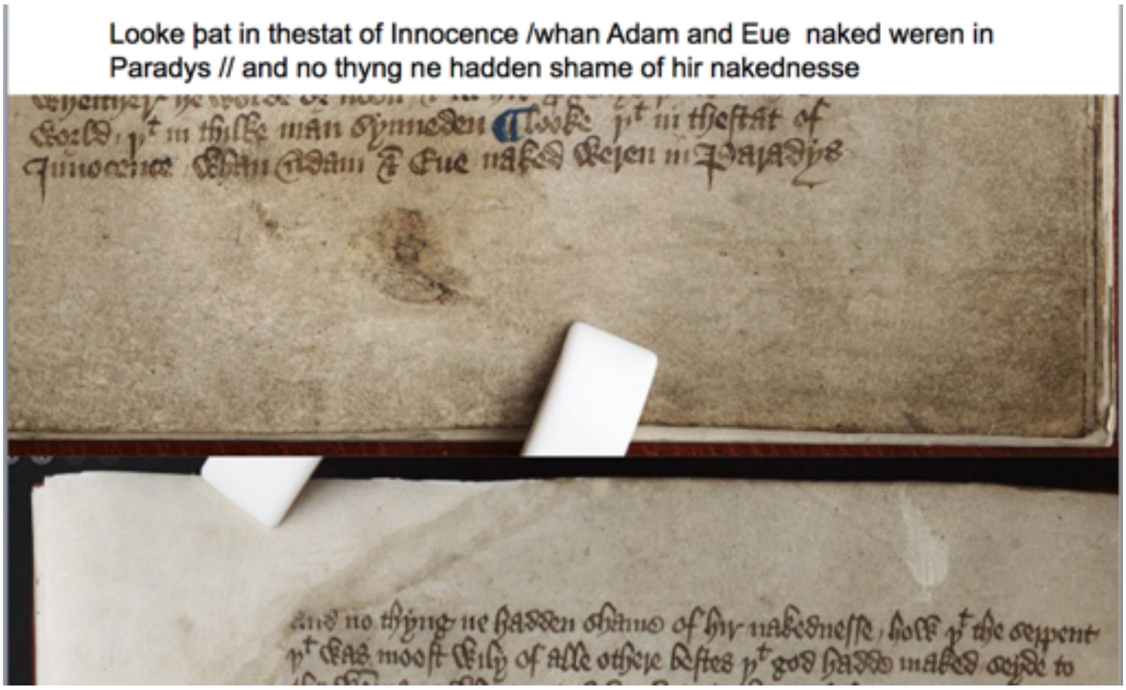
A text is an instance of an act of communication inscribed in a document

Accordingly, every text which ever existed, and ever could exist, has two fundamental characteristics. It is an instance of an act of communication, which may be represented by an ordered hierarchy of content objects. It is also inscribed in a document.

And here is where it gets interesting. A document may also be represented as an ordered hierarchy of content objects. It is made up of quires; the quires are made up of pages; the pages are made up of columns, divided into lines, divided into word spaces. So a document may be represented as a tree, as may too the act of communication.

This is why Durand and his co-authors were exactly half right. They offered a model of the act of communication as a tree. But a text is, as we define it, is not just an act of communication: it is also the document it is inscribed in. A complete representation of a text therefore presents it as two trees: one tree for the act of communication, one tree for the document.

Two trees, not one. Now let's make things even more interesting. Consider paragraph 251 of the Parson's Tale in the Hengwrt manuscript. Read as an act of communication, this is a single statement (or entity, as we call it):

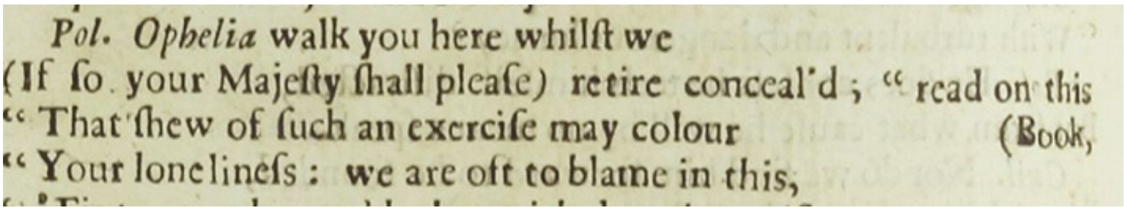


Looke þat in thestat of Innocence /whan Adam and Eue naked weren in
Paradys // and no thyng ne hadden shame of hir nakednesse

Figure 3. Folios 231r and 231v of the Hengwrt manuscript, with Parson's Tale 251

Hence, in the communication act tree: this sentence is on a single branch. But in the Hengwrt manuscript, this sentence is split across two pages, the first half appearing on the base of 231r, the second half on the top of the next page. Accordingly, this appears on two branches in the document tree.

So what, you might say. This is just an overlapping hierarchy. The text is in the same order, just differently divided. Now, look at thjs line:



Pol. Ophelia walk you here whilſt we
(If ſo your Majeſty ſhall pleaſe) retire conceal'd ; " read on this
" That ſhew of ſuch an exerciſe may colour (Book,
" Your lonelineſs : we are oft to blame in this,

Figure 4. From the sixth quarto printing of Hamlet

Here is the well known phenomenon of the turned under line. In this case: the text is not in the same order in the two trees. Indeed, at this point the branches of the two trees are quite different. In the communication tree “read on this book” is in a single branch; in the document tree “on this” and “book” are separated from each other by the text “That shew of such an exercise may colour..” (sixth quarto edition).

If you are still doubtful: look at this example, from the Complutensian polyglot bible.

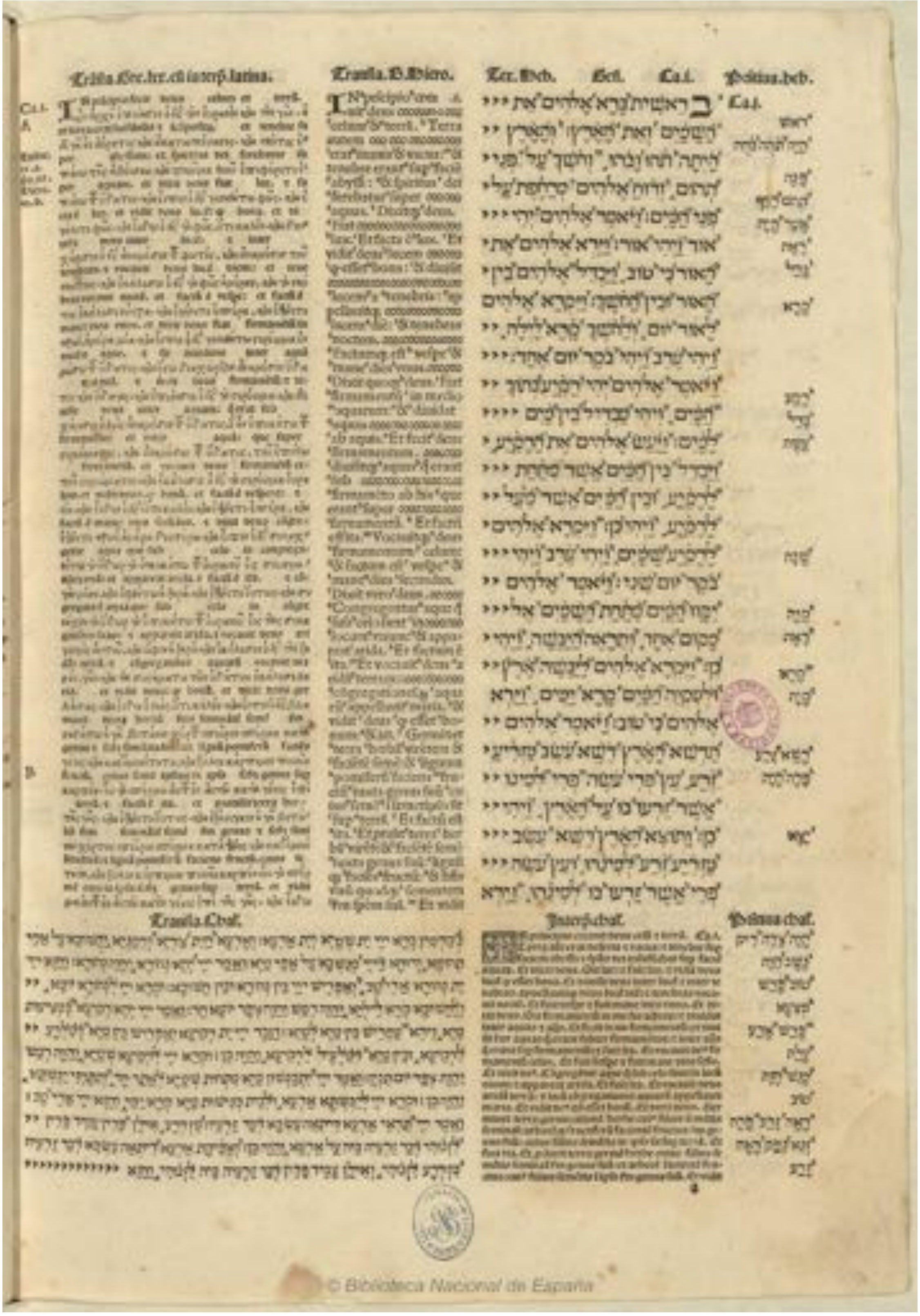
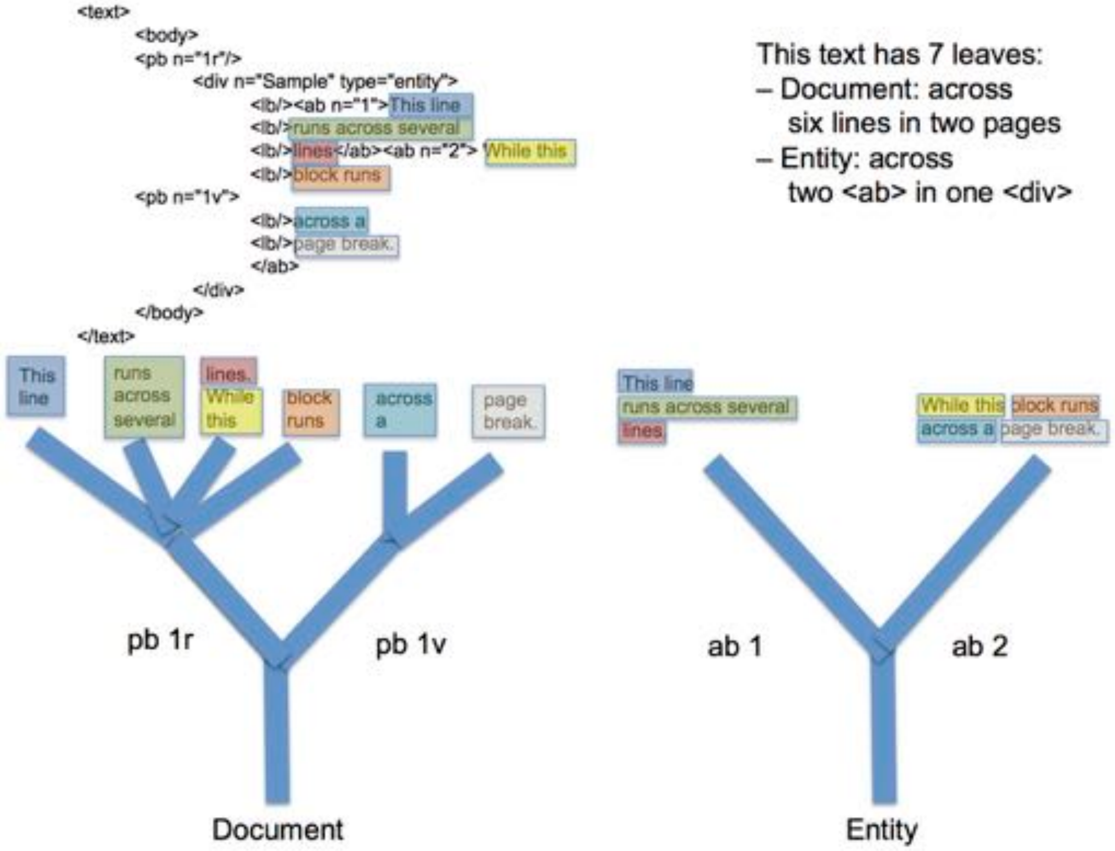


Figure 5. The Complutensian polyglot Bible.

Not only does this single document page contain multiple acts of communication: two of these, the Hebrew on the top right and the Aramaic at the bottom left, are to be read right to left. In these cases, the leaves of text in the document tree and the leaves in the communication act tree will be in reverse order.

In this model, text is a set of leaves appearing on two quite distinct trees. Each tree is completely separate from the others. On one tree, many leaves may hang in one order from a single branch. On the other tree, those leaves may be on many different branches, perhaps widely separated from each other, perhaps in a completely different order. Imagine two trees in nature that somehow have grown into each other, so that although their branches are completely distinct, yet they have the same leaves. Here is how we use this understanding for the simple text we looked at earlier:



```
<text>
        <body>
        <pb n="1r"/>
                <div n="Sample" type="entity">
                        <lb/><ab n="1">This line
                        <lb/>runs across several
                        <lb/>lines</ab><ab n="2"> While this
                        <lb/>block runs
        <pb n="1v">
                        <lb/>across a
                        <lb/>page break.
                        </ab>
                </div>
        </body>
</text>
```

This text has 7 leaves:
– Document: across six lines in two pages
– Entity: across two <ab> in one <div>

Figure 6. A text with seven leaves

We see that the two sentences are decomposed into seven leaves, each holding one or more words of text. These seven leaves are then distributed across two trees. In the document tree, the seven leaves are distributed across six lines in two pages. In the entity tree, the seven are distributed across two <ab> elements in one <div>.

There are many consequences of this definition of text as leaves shared by two trees. First, while each text must appear in two trees, there are many scenarios where there may be more than two: we might want analyze the text rhetorically, or metaphorically, or thematically, each having its own tree. Second, the complete independence of the trees means that questions of how the trees relate to each other have no meaning. Does the communication act precede the page? Or the reverse? Do we have:

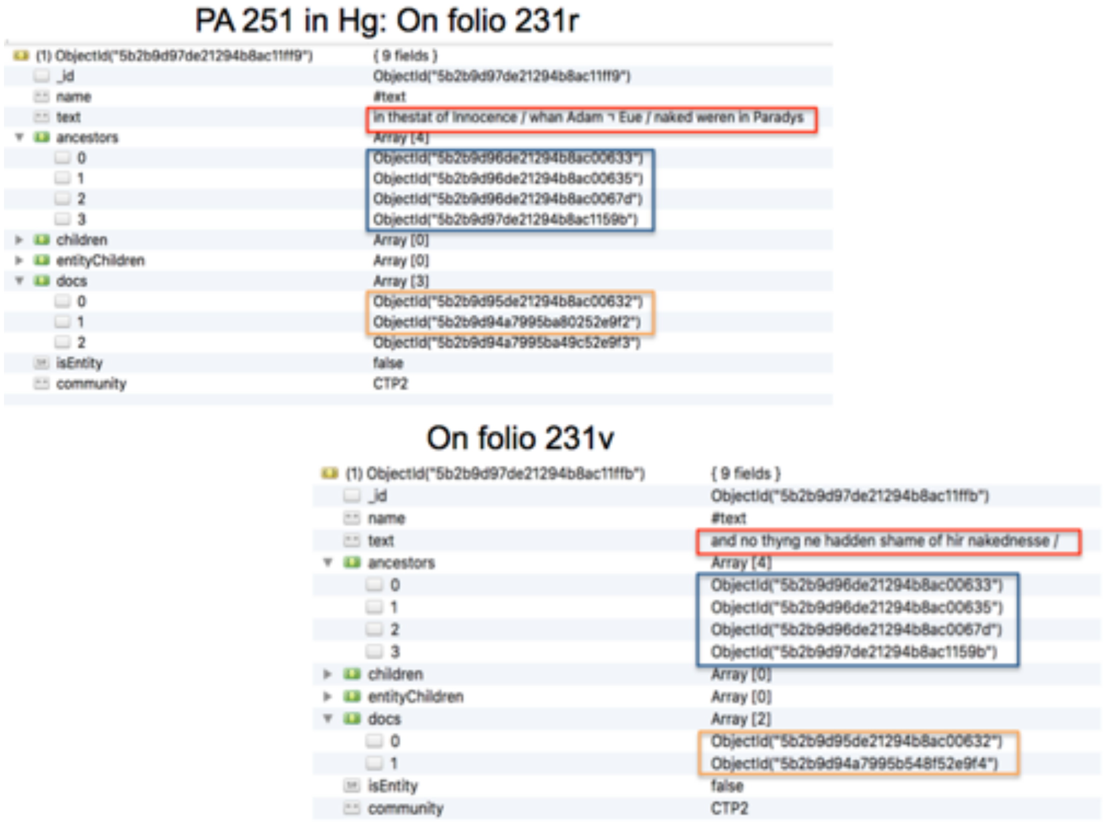
<page><communication act>…
or
<communication act><page>…
It really is irrelevant. The two exist in different dimensions. The text exists in both dimensions: but that is all they share.

In our language, an act of communication is an entity. We represent this formally in this X-Path like syntax:

entity=CT:Part=GP:Line=1 is the first line of the General Prologue of the Canterbury Tales
document=Hg:folio=1r is the first folio of the Hengwrt manuscript
entity=CT:Part=GP:Line=1:document=Hg:folio=1r is the text of the first line of the General Prologue as it appears on the first folio.

So far, the theory of documents, entities and texts behind Textual Communities. What about implementation? A fundamental requirement was that we would allow real-time editing of the two trees and the text they share. You can see the result in Textual Communities.

How did we make this happen? Well, it only took twelve years. There did not seem to be any readymade system doing what we wanted. First, we tried to use an XML database: XML DB, now maintained by Oracle. It took four years to figure out that this would just not fly for us. In 2010, we abandoned XML DB, and spent three years trying to persuade MySQL that it could do what we wanted. But our joins grew so large and complex and slow that in 2013 we abandoned this too. We experimented with SparkL and rdfs: after all, we now had a complex and rich ontology. This too could not work fast enough for what we needed. Finally, we hit upon JSON and MongoDB. And six years later, finally, here we are. For the curious, this is what the core JSON document which holds it all together looks like, for block 251 of the Parson’s Tale, as we saw it earlier:



PA 251 in Hg: On folio 231r

```
(1) ObjectId("5b2b9d97de21294b8ac11ff9")     { 9 fields }
    _id                                      ObjectId("5b2b9d97de21294b8ac11ff9")
    name                                     #text
    text                                     in thestat of Innocence / whan Adam ⁊ Eue / naked weren in Paradys
  ▼ ancestors                                Array [4]
      0                                      ObjectId("5b2b9d96de21294b8ac00633")
      1                                      ObjectId("5b2b9d96de21294b8ac00635")
      2                                      ObjectId("5b2b9d96de21294b8ac0067d")
      3                                      ObjectId("5b2b9d97de21294b8ac1159b")
  ▶ children                                 Array [0]
  ▶ entityChildren                           Array [0]
  ▼ docs                                     Array [3]
      0                                      ObjectId("5b2b9d95de21294b8ac00632")
      1                                      ObjectId("5b2b9d94a7995ba80252e9f2")
      2                                      ObjectId("5b2b9d94a7995ba49c52e9f3")
    isEntity                                 false
    community                                CTP2
```

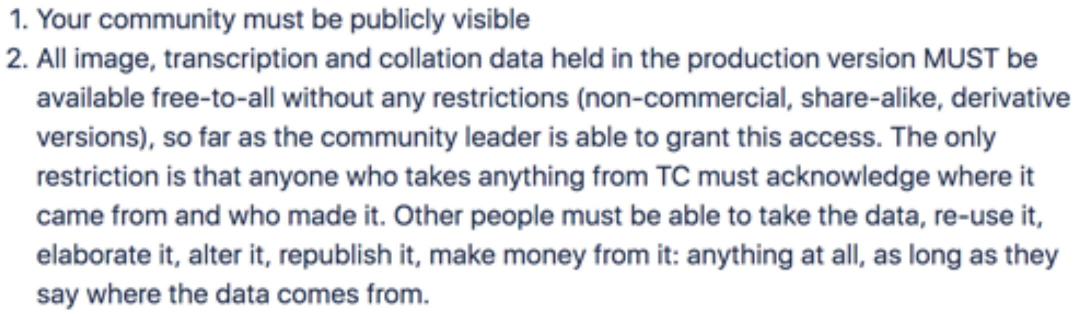
On folio 231v

```
(1) ObjectId("5b2b9d97de21294b8ac11ffb")     { 9 fields }
    _id                                      ObjectId("5b2b9d97de21294b8ac11ffb")
    name                                     #text
    text                                     and no thyng ne hadden shame of hir nakednesse /
  ▼ ancestors                                Array [4]
      0                                      ObjectId("5b2b9d96de21294b8ac00633")
      1                                      ObjectId("5b2b9d96de21294b8ac00635")
      2                                      ObjectId("5b2b9d96de21294b8ac0067d")
      3                                      ObjectId("5b2b9d97de21294b8ac1159b")
  ▶ children                                 Array [0]
  ▶ entityChildren                           Array [0]
  ▼ docs                                     Array [2]
      0                                      ObjectId("5b2b9d95de21294b8ac00632")
      1                                      ObjectId("5b2b9d94a7995b548f52e9f4")
    isEntity                                 false
    community                                CTP2
```

Figure 7. MongoDB documents representing block 251 of the Parson’s Tale

Here, we see the two halves of this block. The entity tree is shown, here, in the ancestors array. Each half of the block is in the same branch of the entity tree: on this tree, they are a single text. But the document tree, shown here in the docs array, is different.

So far, the intellectually original part of the work we did. I said I would speak of the social aspect. You will find two versions of textual communities online: the “production” version and the “sandbox” version. As the name implies: the ‘production’ version offers more support, and a guarantee of data persistence, but only to a few approved projects. There are no such guarantees for the sandbox version: on the other hand, anyone can access it. So, what do you need to do to have your edition on the production version? Here is what we say:



1. Your community must be publicly visible
2. All image, transcription and collation data held in the production version MUST be available free-to-all without any restrictions (non-commercial, share-alike, derivative versions), so far as the community leader is able to grant this access. The only restriction is that anyone who takes anything from TC must acknowledge where it came from and who made it. Other people must be able to take the data, re-use it, elaborate it, alter it, republish it, make money from it: anything at all, as long as they say where the data comes from.

Behind this requirement is a view of how we should conduct ourselves as scholars in the digital age. It is the model of the open source community: what Clay Shirkey calls "design for generosity". In this model, scholarship is not created by scholars working with all the privilege of tenure, publishing in approved journals, rewarded and regulated by learned bodies. It is created by legions: by many people reading, commenting, contributing, giving to each other. 22 years ago, a much younger Michael Sperberg-McQueen was sitting in Newark Airport after a three-day meeting on software tools for humanities scholars. Here is what he wrote about the meeting:

> … the one point on which everyone seems agreed is that we need an open, extensible system, to work with texts we have not read yet, on machines that have not been built yet, performing analyses we have not invented yet. This is not a system for which we can plan the details in advance; its architecture, if we insist on calling it that, will be an emergent property of its development, not an a priori specification. We are not building a building; blueprints will get us nowhere. We are trying to cultivate a coral reef; all we can do is try to give the polyps something to attach themselves to, and watch over their growth.

I think Michael's metaphor applies not just to the digital humanities. This is how scholarship has always proceeded. Each person adds their fragment, as best they can, to the mounting coral reef of knowledge. However, there is a catch. This can only happen if people can take from the reef what they need, make something of it, put it back, and be prepared to let someone else take what they have given. If we do this, then we can make something marvellous.

If ever I am asked "what is the value of the digital humanities" my answer is this: our great merit is that we have a culture of generosity. We give, that others may prosper; and our prosperity in turn depends on others. In thirty years of academic life, I have found nothing so destructive as the instinct of "it's mine": this is my data. I will control who uses it. More than anything else, the aim of Textual Communities is to offer a tool which allows scholars and readers to give to others. If people do this, we can change rather a lot.

Van Gogh Letters: the TEI Publisher clone

Authors: Magdalena Turska, Wolfgang Meier

Important part of work on the TEI Publisher is to test it on new examples and often such tests manifest certain shortcomings or bring ideas for missing functionality. Recent release of the XML sources of Van Gogh Letters under CC BY-NC-SA 4.0 licence gave us an excellent opportunity to try and reconstruct a complete edition – by many regarded as a model digital edition of correspondence – in the newest version of the TEI Publisher.

The TEI Publisher version sports dynamic view panels for different perspectives on the letter - transcription, translation, facsimile and commentaries. It aims to resemble closely the rendering from the original website though it was not the point of the exercise to mimic it exactly and in few cases we decided on alternative rendering. Worth noting is the addition of the faceted search – an upcoming feature in eXist-db 5.0.0 – for which this app was a development test case.

We hope that the TEI Publisher / Van Gogh demo will give the opportunity to show at a glance main new and old features of the TEI Publisher and point out some of the challenges it tries to address as a general purpose publishing platform.

PS Please note that the TEI Publisher edition of Van Gogh letters is just a demo showcasing eXist-db and TEI Publisher and does not intend to compete with or supplement the canonical version in any way.

TE
2019