The Linked TEI: Text Encoding in the Web

TEI Conference and Members Meeting 2013

Book of Abstracts

edited by

Fabio Ciotti

Arianna Ciula

ASSOCIAZIONE PER L'INFORMATICA UMANISTICA E LA CULTURA DIGITALE

DIGILAB
CENTRO INTERDIPARTIMENTALE
DI RICERCA E SERVIZI

SAPIENZA
UNIVERSITÀ DI ROMA

The Linked TEI: Text Encoding in the Web

Book of Abstracts - electronic edition

Abstracts of the TEI Conference and Members Meeting 2013:
October 2-5, Rome

Edited by Fabio Ciotti and Arianna Ciula

DIGILAB Sapienza University & TEI Consortium
Rome 2013

Contents

Introduction

Ciotti, Fabio; Ciula, Arianna

This year's conference focuses on the concept of linked text encoding, encouraging reflections on the semantics of the TEI conceptual model, but also placing the TEI within a framework of interconnected digital resources.

The title we chose, "The Linked TEI: Text Encoding in the Web", hints obviously at a very actual theme in the digital realm: the emergence and diffusion of the Linked Data paradigm and of a Participatory Web. TEI has had a crucial - and nowadays widely recognised - role in encouraging and facilitating the creation of vast amounts of textual and linguistic resources. The TEI has been one of the major enabling technologies in the Digital Humanities. However, the dominant paradigm in the creation of digital resources, especially in the academic domain, has been that one of the monad archive, of the big or small project, *perfectum* in itself. We think that to continue in its role of stimulus for innovation in the Digital Humanities, TEI has to be able to embrace fully the new paradigm. Of this paradigm, the sharing and the interconnection of data on the Web as well as the emergence of the semantic level of data are the most interesting aspects able to bring about new developments. But in the idea of "Linked TEI" the issues around multilingualism and multiculturalism are also encompassed: to be connected means to to be able to adapt to different traditions and languages.

Contributions have responded very well to the challenge with a rich range of topics and perspectives being represented in the programme: from reflections on semantic models and texts as such, to data modelling and tools for analysis, from re-thinking research infrastructures and developing participatory approaches, to establishing mutual linking between dictionaries and corpora. The precious expertise of 34 reviewers based in 10 different countries allowed the programme committee (composed of Arianna Ciula, Lou Burnard, Marjorie Burghart, Sebastian Rahtz, Gianfranco Crupi and Fabio Ciotti) to craft a menu out of an interesting combination of traditional and innovative ingredients.

The community has grown fast over the years and in multiple directions. The programme clearly reveals this too. Indeed, with the plus that Rome is an easy (and pleasant) location to reach - special thanks to the *Associazione Italiana di Informatica Umanistica e Cultura Digitale* (AIUCD), and to Digilab Sapienza (an interdepartmental center whose mission is to promote the interdisciplinarity in humanities research and to communicate and promote the cultural heritage in the digital environment), that carried most of the organizational effort and logistic support. We wish also remember our sponsors: ICCU (Central Institute for the Union Catalogue of Italian Libraries), CINECA, BUCAP, Synchro Soft, EADH and Google for making the conference in this venue possible - participation was particularly high (at the time of writing: more than 150 attendees based in countries all over the world) and rich in representation of different languages of interest (Arabic, Armenian, Berber, medieval Castilian, Dutch, Egyptian, English, French, Georgian, historical and modern German, ancient Greek, Italian, Japanese, classic and medieval Latin, Portuguese, Spanish, Syriac), writing systems (including Hieroglyphic and Tifinagh) and genres (from librettos to notebooks, from 15th century *Cancioneros* to social media chats).

It is of particular honour to us as Italians that this conferences took place in Rome and at the University of La Sapienza. Our country has given high contributions to the history of TEI and Digital Humanities as a whole in terms of theory, technologies and organisation: it suffices to mention the name of the late Antonio Zampolli, who contributed enormously to the conception and implementation of the TEI project and who hosted the first Members' Meeting of the consortium in Pisa in 2001. The group of scholars, mostly young, who met in this University under the mentorship of Giuseppe Gigliozzi has played an important role too. It was thanks to his intelligence that this group perceived the importance of a formal instrument for the representation of texts such as the TEI. Since the mid-90s they have been digitizing and encoding texts, first sporadically, then in a structured manner in the context of the TIL (Italian Texts Online) project funded by the Ministry of University. We can say without fear of contradiction that, thanks to these efforts, the TEI has spread throughout Italy and that these works and associated teaching activities have trained

many and many young scholars, some of whom have then gone on to a brilliant scientific career, both nationally and internationally.
Every now and then it is healthy to ask where the TEI is going and why we care about it. With the growing emergence of Digital Humanities curricula and research positions, the establishment of digital workflows and resources in the cultural heritage sector, the convergence towards a digital scholarly communication cycle, it is easy to get caught in the vortex. In all this, is the TEI surpassed? A step back to think of where we are - and therefore where we are going - leads us to the historical focus of the TEI: texts, mainly its modelling for processing purposes. We believe that it is indeed its historical focus together with a periodical shifting of its limits and limitations that makes the TEI actual: the meanings of texts change; the term text itself is perceived and applied in a wide sense; what we want to do with texts change. The heart of the problem stays: the slippery confines of our cultural productions, there to be seen, analysed, reflected upon, deconstructed, formalised, processed, remediated and re-interpreted. A glance at the TEI Special Interest Groups - that thanks to the enthusiasm of their conveners met numerous in Rome - range of scope and focus is proof of this if we needed one.
The key to ensure the TEI will have a future in its contribution to scholarship and culture in general lies therefore also in a slight but crucial tilt to its rhetorics: the TEI is not about delivering a standard but rather about creating it in partnership with the diverse communities of researchers, archivists, librarians and other professionals of the cultural heritage sector, software developers, infrastructures providers, artists, citizens.
The pictures of a linked TEI looks less and less like a jigsaw where all pieces are cut to fit together, and more like an intertwining of hands.
Enjoy the reading.

Keynote speeches

Faceting Text Corpora

Demonet, Marie Luce

The BVH (Bibliothèques Virtuelles Humanistes) team of the CESR-University of Tours started using the TEI encoding scheme to annotate and publish French Renaissance texts in 2006. In 2011, the "Corpus" research network was set up with an aim to developing the field of digital humanities and the European research infrastructures roadmap. Supported by the French Ministry of Research, this network comprises several national consortia, amongst which "CAHIER" (Corpus d'Auteurs pour les Humanités: Informatisation, Edition, Recherche – Authors' Corpora for the Humanities: Digitisation, Edition, Research), coordinated by the CESR. The corpora assembled by this consortium (over 25) mainly centre on literary figures, but also concern the work of philosophers and the history of science: Polish philosophers, d'Alembert, Machiavel, Montaigne, Flaubert, Montesquieu, etc. CAHIER regularly organises training sessions and workshops aimed at developing linguistic, thematic and philological approaches to online editing using TEI guidelines. It is also involved in a number of workshops such as management projects and tutorials on specific tools (CMS, OAI-PMH, TXM, PhiloLogic). Collaboration with the two linguistic consortia (oral and written corpora) of the "Corpus" research network is already underway: a jointly organised advanced TEI workshop was programmed and a joint reflection on the definition(s) of "corpora" engaged. A special interest group, bringing together linguists, medievalists and anthropologists, is working on a series of recommendations concerning copyright. Thanks to long term collaborations with numerous libraries (French National Library, university and public libraries, Europeana Libraries consortium), set up well before the creation of the "Corpus" network, we are able to benefit from their staff's expertise in metadata, records, iconographic thesauri and bibliographical databases. The aim of the CAHIER consortium is not just to provide online facsimiles of the assembled corpora, but to offer full-text documents, searchable, retrievable and shareable in XML and standard formats.

Online editing of collections of fragments is a preoccupation shared by many a scholar. A special interest group dedicated to correspondences is currently being set up which will be able to use TEI schemas already being exploited by several projects. What I particularly wish to bring attention to though is the potential added-value of the TEI guidelines in the domain of corpora publication as a means of furthering typological and taxonomic approaches to text processing.

Many TEI guideline users (and non-users) regret the lack of search tools and browsers capable of obtaining relevant results through a "genre tree". At present, libraries all have their own thesauri, generally unknown to scholars; no initiatives seem to exist with an aim to developing folksonomies for texts, nor does there appear to be any coordination between libraries and booksellers, who have their own way of classifying their products. The situation is not however stalemate: one relevant starting point for moving forward would, for example, be to combine the dichotomy between fiction and non-fiction with the formal schemas found in the TEI guidelines: prose, verse and drama. A new schema would not be needed; only a well organised thesaurus for genres and sub-genres embedded using TEI and RDF; this could be searched through the headers using a faceted browser and adapted to a wide range of languages. Two objections obviously spring to mind. The first is the difference between national traditions: providing a multilingual tool would in itself necessitate a major research project. The second is the difficulty scholars have in agreeing on the definition of genres and their ontologies. I fully acknowledge that the development of an interdisciplinary and interlinguistic thesaurus presents a considerable challenge, but seems to me that it is well worth rising to.

Biography

Marie-Luce Demonet is professor of French Renaissance literature and director of the Maison des Sciences de l'Homme Val de Loire (The Loire Valley Institute for Social Sciences and Humanities).

Specialist of the relationship between literature and language, Marie-Luce Demonet has written works on relevant French authors and humanists such as Rabelais, Montaigne and Pasquier (critical and electronic editions, conference proceedings, monographs), and on the issues of literary

theory (novel, fiction) and semiotics. Mrs. Demonet is the creator of two websites which host original texts of Renaissance (e.g. http://www.bvh.univ-tours.fr/Epistemon) and head of the project "Bibliothèques virtuelles des humanistes" (Humanists' Virtual Libraries). Furthermore, she has published several articles concerning the application of the new technologies to the French Renaissance literature and taken part, since 1990, in various events about the same topic.
Her main areas of research include:

- History of Linguistic Theories
- Literary Genres
- Electronic Editions
- Philosophy of Language and Literature

Recommended publications are:

- *Michel de Montaigne, Les Essais*. Marie-Luce Demonet. Presses Universitaires de France (2002).
- *Montaigne et la Question de l'Homme*. Marie-Luce Demonet. Presses Universitaires de France (1999).
- *Les Voix du Signe: Nature et Origine du Langage à la Renaissance, 1480-1580*. Marie-Luce Demonet. H. Champion (1992).
- *Les Grands Jours de Rabelais en Poitou: Actes du colloque international de Poitiers des 30 août et 1er septembre 2001*. Marie-Luce Demonet. Droz (2006).

Text encoding, ontologies, and the future

Renear, Allen

SGML/XML text encoding has played a important role in the development of the global networked information system that now dominates almost all aspects of our daily lives — commercial, scientific, political,

social, cultural. The TEI community in particular has made impressive contributions. Today the information organization strategies that provide the foundation for contemporary information technologies are undergoing a new phase of intense and ambitious development. There has of course been a period of skepticism, just as there was with SGML in the 1980s. But that period is now behind us, or should be. Ontologies, “linked open data”, and semantic web languages like OWL and RDF have proven their value and are beginning to yield practical applications. These developments are not radical new strategies in information organization, rather they are continuation of a long-standing trajectory towards increased abstraction, declarative formalization, and standardization — strategies with a solid track record of success. In the last thirty years the text encoding community has helped sustain and advance the evolution of these information organization strategies, and is now well-positioned to further contribute to, and exploit, recent developments.

I will discuss the significance of all this not only for libraries, publishing, data curation, and the digital humanities, but also for the global networked information system more generally. Without a doubt advances in formalization will continue to bring us many new advantages, and so there is much to look forward to. But at the same time the low-hanging fruit has been picked and the problems we will encounter in the next decade or two will prove quite challenging.

Biography

Allen Renear is professor and interim Dean at GSLIS (the Graduate School of Library and Information Science, University of Illinois, USA) where he teaches courses and leads research in information modeling, data curation, and digital publishing. Prior to coming to GSLIS Allen Renear was the Director of the Brown University Scholarly Technology Group. He received an AB from Bowdoin College and an MA and PhD from Brown University.

Recently, Renear's work has focused on fundamental issues in the curation of scientific datasets and conceptual models for data management and preservation. This includes topics such as levels of abstraction and encoding, identity, ontology, etc., as well as projects in several related areas:

- A Formal Framework for Data Concepts
- Ontologies to Support *Strategic Reading*
- Collection/Item Metadata Relationships
- Ontologies for Digital Objects

The projects are all affiliated with the GSLIS Center for Informatics Research in Science and Scholarship and funded by the National Science Foundation, the Institute for Museum and Library Studies, and the Mellon Foundation.
Recommended recent pubblications are:

- *Strategic Reading, Ontologies, and the Future of Scientific Publishing.* Allen H. Renear, Carole L. Palmer. Science. 325:5942 p. 828 (2009).
- *When Digital Objects Change — Exactly What Changes?*. Allen H. Renear, David Dubin, Karen M. Wickett. Proceedings of the American Society for Information Science and Technology. 45:1 (2008).

Further works selected by Allen Renear for the interested can be found here http://people.lis.illinois.edu/~renear/renearcv.html, ordered in three categories:

- Ontology of Scientific and Cultural Objects
- Metadata and Logic
- Semantic Approaches to Digital Publishing

Papers

The Linked Fragment: TEI and the encoding of text re-uses of lost authors

Berti, Monica; Almas, Bridget

The goal of this paper is to present characteristics and requirements for encoding quotations and text re-uses of lost works (i.e., those pieces of information about lost authors that humanists classify as 'fragments'). In particular the discussion will focus on the work currently done using components of Perseids (http://sites.tufts.edu/perseids/), a collaborative platform being developed by the Perseus Project that leverages and extends pre-existing open-source tools and services to support editing and annotating TEI XML source documents in Classics.

Working with text re-uses of fragmentary authors means annotating information pertaining to lost works that is embedded in surviving texts. These fragments of information derive from a great variety of text re-uses that range from verbatim quotations to vague allusions and translations. One of the main challenges when looking for traces of lost works is the reconstruction of the complex relationship between the text re-use and its embedding context. Pursuing this goal means dealing with three main tasks: 1) weighing the level of interference played by the author who has reused and transformed the original context of the information; 2) measuring the distance between the source text and the derived text; 3) trying to perceive the degree of text re-use and its effects on the final text.

The first step for rethinking the significance of quotations and text re-uses of lost works is to represent them inside their preserving context. This means first of all to select the string of words that belong to the portion of text which is classifiable as re-use and secondly to encode all those elements that signal the presence of the text re-use (i.e., named entities such as the onomastics of re-used authors, titles of re-used works and descriptions of their content, *verba dicendi*, syntax, etc.). The second step is to align and encode all information pertaining to other sources that reuse the same original text with different words or a different syntax (witnesses), or that deal with the same topic of the text re-use (parallel

texts), and finally different editions and translations of both the source and the derived texts.

This paper addresses the following requirements for producing a dynamic representation of quotations and text re-uses of fragmentary authors, which involve different technologies including both inline and stand-off markup:

- *Identifiers*: i.e. stable ways for identifying: fragmentary authors; different kinds of quotations and text re-uses; passages and works that preserve quotations and text re-uses; editions and translations of source texts; entities mentioned within the text re-uses; annotations on the text re-uses.
- *Links*: between the fragment identifier and the instances of text re-use, the fragment identifier and the attributed author, the fragment identifier and an edition which collects it; between the quoted passage and the entities referenced in it; between the quoted passage and translations.
- *Annotations*: the type of re-use; canonical citations of text re-uses; dates of the initial creation of the re-use, of the work which quotes it, author birth and death; editorial commentary on each text re-use; bibliography; morphosyntactic analysis of the quoted passage; text re-use analysis (across different re-uses of the same text); syntactic re-use analysis; translation alignments (between re-used passages and their translations); text reuse alignments (between different re-uses of the passage in the same language).
- *Collections* (the goal is to organize text re-uses into the following types of collections): all text re-uses represented in a given edition which includes re-uses from one or many authors; all text re-uses attributed to a specific author; all text re-uses quoted by a specific author; all text re-uses referencing a specific topic; all text re-uses attributed to a specific time period, etc.

In the paper we discuss in particular how we are combining TEI (http://www.tei-c.org), the Open Annotation Collaboration (OAC) core data model (http://www.openannotation.org/spec/core/), and the CITE Architecture (http://www.homermultitext.org/hmt-doc/cite/index.html) to

represent quotations and text re-uses via RDF triples. The subject and object resources of these triples can be resolved by Canonical Text and CITE Collection Services to supply the TEI XML and other source data in real time in order to produce new dynamic, data-driven representations of the aggregated information.

The CITE Architecture defines CTS URNs for creating semantically meaningful unique identifiers for texts, and passages within a text. It also defines an alternate identifier syntax, in the form of a CITE URN, for data objects which don't meet the characteristics of citable text nodes, such as images, text re-uses of lost works, and annotations. As URNs, these identifiers are not web-resolvable on their own, but by combining them with a URI prefix and deploying CTS and CITE services to serve the identified resources at those addresses, we have resolvable, stable identifiers for our texts, data objects and annotations. In the paper we supply specific examples of URNs, and their corresponding URIs, for texts, citations, images and annotations.

The CTS API for passage retrieval depends upon the availability of well-formed XML from which citable passages of texts can be retrieved by XPath. The TEI standard provides the markup syntax and vocabulary needed to produce XML which meets these requirements, and is a well-accepted standard for digitization of texts. Particularly applicable are the TEI elements for representing the hierarchy of citable nodes in a text. The Open Annotation Core data model provides with us a controlled vocabulary to identify the motivation for the annotations and enables us to express our annotation triples according to a defined and documented standard.

In the paper we present practical examples of annotations of text re-uses of lost works that have been realized using components of the Perseids platform. In Perseids we are combining and extending a variety of open source tools and frameworks that have been developed by members of the Digital Classics communitity in order to provide a collaborative environment for editing, annotating and publishing digital editions and annotations. The two most prominent components of this platform are the Son of SUDA Online tool developed by the Papyri.info (http://papyri.info) project and the CITE architecture, as previously mentioned.

The outcome of this work is presented in a demonstration interface of Perseids, The Fragmentary Texts Demo (http://services.perseus.tufts.edu/berti_demo/). We also present the data driving the demo, which contains sets of OAC annotations (http://services.perseus.tufts.edu/berti_demo/berti_annotate.js) serialized according to the JSON-LD specification.
The final goal is to publish the annotations and include all the information pertaining to fragmentary texts in the collection of Greek and Roman materials in the Perseus Digital Library. The purpose is to collect different kinds of annotations of text re-uses of fragmentary authors with a twofold perspective: 1) going beyond the limits of print culture collections where text re-uses are reproduced as decontextualized extracts from many different sources, and representing them inside their texts of transmission and therefore as contextualized annotations about lost works; 2) allowing the user to retrieve multiple search results using different criteria: collections of fragmentary authors and works, morphosyntactic data concerning text re-uses, information about the lexicon of re-used words, cross-genre re-uses, text re-use topics, etc.

Bibliography

- Almas, Bridget and Beaulieu, Marie-Claire (2013): *Developing a New Integrated Editing Platform for Source Documents in Classics*. In: Literary and Linguistic Computing (Digital Humanities 2012 Proceedings) (forthcoming).
- Berti, Monica (2013): *Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres*. In: Ancient Society 43.
- Berti, Monica, Romanello, Matteo, Babeu, Alison and Crane, Gregory R. (2009): *Collecting Fragmentary Authors in a Digital Library*. In: Proceedings of the 2009 Joint International Conference on Digital Libraries (JCDL '09). Austin, TX. New York, NY: ACM Digital Library, 259-262.
  http://dl.acm.org/citation.cfm?id=1555442
- Büchler, Marco, Geßner, Annette, Berti, Monica, and Eckart, Thomas (2012): *Measuring the Influence of a Work by Text Reuse*. In: Dunn, Stuart and Mahony, Simon (Ed.): Digital Classicist

Supplement. Bulletin of the Institute of Classical Studies. Wiley-Blackwell.

- Crane, Gregory R. (2011): *From Subjects to Citizens in a Global Republic of Letters*. In: Grandin, Karl (Ed.): Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Nobel Symposium 147. The Nobel Foundation, 251-254.
- Romanello, Matteo, Berti, Monica, Boschetti, Federico, Babeu, Alison and Crane, Gregory R. (2009):*Rethinking Critical Editions of Fragmentary Texts by Ontologies*. In: ELPUB 2009: 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies. Milan, 155-174. http://hdl.handle.net/10427/70403
- Smith, D. Neel and Blackwell, Chris (2012): *Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture*. In: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends. The Center of Hellenic Studies of Harvard University. http://folio.furman.edu/projects/cite/four_urls.html

"Reports of My Death Are Greatly Exaggerated": Findings from the TEI in Libraries Survey

Dalmau, Michelle; Hawkins, Kevin S.

Historically libraries, especially academic libraries, have contributed to the development of the TEI Guidelines, largely in response to mandates to provide access to and preserve electronic texts (Engle 1998; Friedland 1997; Giesecke, McNeil, and Minks 2000; Nellhaus 2011). At the turn of the 21st century, momentum for text encoding grew in libraries as a result of the maturation of pioneering digital library programs and XML-

based web publishing tools and systems (Bradley 2004). Libraries were not only providing "access to original source material, contextualization, and commentaries, but they also provide[ed] a set of additional resources and service[s]" equally rooted in robust technical infrastructure and noble "ethical traditions" that have critically shaped humanities pedagogy and research (Besser 2004).

In 2002, Sukovic posited that libraries' changing roles would and could positively impact publishing and academic research by leveraging both standards such as the TEI Guidelines and traditional library expertise, namely in cataloging units due to their specialized knowledge in authority control, subject analysis, and of course, bibliographic description. Not long after, in 2004, Google announced the scanning of books in major academic libraries to be included in Google Books (Google 2012), and in 2008 many of these libraries formed HathiTrust to provide access to facsimile page images created through mass digitization efforts (Wilkin 2011), calling into question the role for libraries in text encoding that Sukovic advocated. In 2011, with the formation of the HathiTrust Research Center and IMLS funding of TAPAS (TEI Archiving, Publishing, and Access Service, http://www.tapasproject.org/), we see that both large- and small-scale textual analysis are equally viable and worthy pursuits for digital research inquiry in which libraries are heavily vested (Jockers and Flanders 2013). More recently, we are witnessing a call for greater and more formal involvement of libraries in digital humanities endeavors and partnerships (Vandegrift 2012; Muñoz 2012) in which the resurgence of TEI in libraries is becoming apparent (Green 2013; Milewicz 2012; Tomasek 2011; Dalmau and Courtney 2011). How has advocating for such wide-ranging library objectives — from digital access and preservation to digital literacy and scholarship, from supporting non-expressive/non-consumptive research practices to research practices rooted in the markup itself — informed the evolution or devolution of text encoding projects in libraries?

Inspired by the papers, presentations and discussions that resulted from the theme of the 2009 Conference and Members' Meeting of the TEI Consortium, "Text Encoding in the Era of Mass Digitization," the launch of the AccessTEI program in 2010, and the release of the Best Practices

for TEI in Libraries in 2011, we surveyed employees of libraries around the world between November 2012 and January 2013 to learn more about text encoding practices and gauge current attitudes about text encoding in libraries. As library services evolve to promote varied modes of scholarly communications and accompanying services, and digital library initiatives become more widespread and increasingly decentralized, how is text encoding situated in these new or expanding areas? Do we see trends in uptake or downsizing of text encoding initiatives in smaller or larger academic institutions? How does administrative support or lack thereof impact the level of interest and engagement in TEI-based projects across the library as whole? What is the nature of library-led or -partnered electronic text projects, and is there an increase or decrease in local mass digitization or scholarly encoding initiatives? Our survey findings provide, if not answers to these, glimpses of the TEI landscape in libraries today.

The survey closed on January 31, 2013, with a total of 138 responses, and a completion rate of 65.2%. Since the survey was targeted specifically toward librarians and library staff, we turned away respondents for not meeting that criterion, with a final total of 90 responses. Most of the respondents are from North America (87%), and affiliated with an academic library (82%). Respondents from academic institutions come from institutions of various sizes, with a plurality (31%) falling in the middle range (10,000-25,000 student enrollment). Of those responding, 81.2% are actively engaged in text encoding projects. Preliminary data analysis shows that those not yet engaged in text encoding (or not sure whether their institution is engaged) are planning to embark on text encoding based on grant funding or new administrative support for text encoding projects. It seems that reports of the death of TEI in libraries are greatly exaggerated, though this is not to say that TEI in libraries is not struggling.

Our paper will unveil a fuller analysis of the data we have gathered, and when applicable, a comparative examination against the following raw data sources and publications for a more complete picture:

- TEI-C membership profile of library institutions from 2005 to 2012

- Evolution/devolution of electronic text centers within libraries from as early as 2000 to present
- Findings from a study by Harriett Green (2012) on library support for the TEI
- Findings from a study by Siemens et al. (2011) on membership and recruitment for the TEI Consortium

Emerging trends and issues will inform the future direction and agenda of the TEI's Special Interest Group on Libraries.

Bibliography

- Besser, Howard., 2004. "The Past, Present, and Future of Digital Libraries." A Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion/.
- Bradley, John. 2004. "Text Tools." A Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion/.
- Dalmau, Michelle and Angela Courtney. 2011. "The Victorian Women Writers Project Resurrected: A Case Study in Sustainability." Paper presented at Digital Humanities 2011: Big Tent Humanities, Palo Alto, California, June 19–22.
- Engle. Michael. 1998. "The social position of electronic text centers." Library Hi Tech 16 (3/4): 15–20. http://dx.doi.org/10.1108/07378839810304522.
- Friedland, LeeEllen. 1997. "Do Digital Libraries Need the TEI? A View from the Trenches." Paper presented at TEI10: The Text Encoding Initiative Tenth Anniversary User Conference, Providence, Rhode Island, November 14–16. http://www.stg.brown.edu/conferences/tei10/tei10.papers/friedland.html.
- Giesecke, Joan, Beth McNeil, and Gina L. B. Minks. 2000. "Electronic Text Centers: Creating Research Collections on a Limited Budget: The Nebraska Experience." Journal of Library

Administration 31 (2): 77–92. http://digitalcommons.unl.edu/libraryscience/63/.

- Google. 2012. "Google Books History." Last modified December 21. http://www.google.com/googlebooks/about/history.html.
- Green, Harriett. 2012. "Library Support for the TEI: Tutorials, Teaching, and Tools." Paper presented at TEI and the C(r l)o(w u)d: 2012 Annual Conference and Members' Meeting of the TEI Consortium, College Station, Texas, November 8–10.
- Green, Harriett. 2013. "TEI and Libraries: New Avenues for Digital Literacy?" dh+lib: Where Digital Humanities and Librarianship Meet. http://acrl.ala.org/dh/2013/01/22/tei-and-libraries-new-avenues-for-digital-literacy/.
- Jockers, Matthew L. and Julia Flanders. 2013. "A Matter of Scale." Keynote lecture presented at Boston-Area Days of DH 2013. http://digitalcommons.unl.edu/englishfacpubs/106/.
- Milewicz, Liz. 2012. "Why TEI? Text > Data Thursday." Duke University Libraries News, Events, and Exhibits. http://blogs.library.duke.edu/blog/2012/09/26/why-tei-text-data-thursday/.
- Muñoz, Trevor. 2012. "Digital Humanities in the Libraries Isn't a Service." Notebook. http://trevormunoz.com/notebook/2012/08/19/doing-dh-in-the-library.html.
- Nellhaus, Tobin. 2001. "XML, TEI, and Digital Libraries in the Humanities." Libraries and the Academy 1(3): 257–77. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v001/1.3nellhaus.html.
- Siemens, Ray, Hefeng (Eddie) Wen, Cara Leitch, Dot Porter, Liam Sherriff, Karin Armstrong, and Melanie Chernyk. 2011. "The Apex of Hipster XML GeekDOM" Journal of the Text Encoding Initiative 1. http://jtei.revues.org/210.
- Sukovic, Suzana. 2002. "Beyond the Scriptorium: The Role of the Library in Text Encoding." D-Lib Magazine 8.1. http://www.dlib.org/dlib/january02/sukovic/01sukovic.html.
- Tomasek, Kathryn. 2011. "Digital Humanities, Libraries, and Scholarly Communication." Doing History

Digitally. http://kathryntomasek.wordpress.com/2011/11/02/digital-humanities-libraries-and-scholarly-communication/.
- Vandegrift, Micah. 2012. "What is Digital Humanities and What's It Doing in the Library?" In the Library with the Lead Pipe. http://www.inthelibrarywiththeleadpipe.org/2012/dhandthelib/.
- Wilkin, John. 2011. "HathiTrust's Past, Present, and Future." Remarks presented at the HathiTrust Constitutional Convention, Washington, D.C., October 8. http://www.hathitrust.org/blogs/perspectives-from-hathitrust/hathitrusts-past-present-and-future.

From entity description to semantic analysis: The case of Theodor Fontane's notebooks

de la Iglesia, Martin; Göbel, Mathias

Within the last decades, TEI has become a major instrument for philologists in the digital age, particularly since the recent incorporation of a set of mechanisms to facilitate the encoding of genetic editions. Editions use the XML syntax while aiming to preserve the quantity and quality of old books and manuscripts, and to publish many more of them online mostly under free licences. Scholars all over the world are now able to use huge data sets for further research. There are many digital editions available, but only a few frameworks to analyse them. Our presentation focusses on the use of web technologies (XML and related technologies as well as JavaScript) to enrich the forthcoming edition of Theodor Fontane's notebooks with a data driven visualisation of named entities and to build applications using such visualisations which are reusable for any other edition within the world of TEI.

State of the art

The TEI Guidelines provide various mechanisms for tagging references to entities in texts, as well as solutions for encoding metadata supplied

by editors about such entities. Such methods are frequently employed in digital editions. For example, on the website of the edition of John Godwin's diaries[1] we are able to highlight the names within the text in different colors. Often these parts are rendered in HTML as <acronym> and are equipped with a box containing further information that pops up as the user clicks on or hovers over them. This is a simple and easy to use way to deliver further information and some search options, but it does not *per se* facilitate a detailed analysis.

With help of the <speaker> tag within TEI encoded drama, a quantitative analysis of spoken words becomes possible. One example is provided by the Women Writers Project, that visualize speakers in drama by gender.[2] It is also possible to get a quantitative overview of the coappearance of two or more characters, which is done for Victor Hugo's Les Misérables with the help of the D3.js JavaScript library.[3]

Persons and places seem to be the most common types of tagged entities. These are usually normalized, i.e. spelling variations are merged and matched to an authoritative name, and some additional data not found in the encoded source text is provided – most commonly biographical dates for persons and geographic coordinates for places. Additional data might include excerpts from encyclopedias, or map visualisations of the location of places. In the case of most editions, the usage of entity encoding can be characterised as descriptive, rather than analytical: information is provided about entities, but the way in which they are referenced in source texts and how the entities relate to each other is recorded and used for navigational purposes only. This paper, employing the example of a TEI edition project of 19th century notebooks, discusses further potential uses of such TEI encoded semantic annotations.

Theodor Fontane's notebooks

From 1859 until the late 1880s, the German poet Theodor Fontane (1819–1898) filled almost 10,000 pages in 67 notebooks, which have not yet been published in their entirety. They include diary entries, travel notes, theater criticism and drafts for novels and poems, resulting in a wide spectrum of text types and images.# The complete edition of the notebooks both in print and online is being prepaired at the Theodor Fontane-Arbeitsstelle, Department of German Philology at Göttingen University,

in collaboration with the Göttingen State and University Library.# In his notebooks, Fontane made extensive use of underlining, cancellations, corrections and additions, and consequently the crucial aspect of the philological edition project is to precisely transcribe, encode, annotate and visualize the appearance of Fontane's handwriting, in order to help the reader to decipher and understand it. Another important task within this project, however, is to identify and encode references to entities in the notebooks. These include:

- persons, organizations – linked to authority files such as GND# or VIAF#, online historical encyclopedias
- places – all of the above, plus linked to geographical databases such as GeoNames or the Getty Thesaurus of Geographic Names
- dates – normalized to machine-readable standards, so that dates can be sorted and durations calculated
- artworks, buildings – linked to their creators, locations, and provided with their dates of creation
- literary works, musical works – linked to their authors and, where applicable, online versions
- events (e.g. battles) – linked to places and provided with dates
- characters in works of fiction – linked to the respective works.

Because of the density of occurrences and the variety of entity types, Fontane's notebooks lend themselves to advanced methods of semantic analysis.

Semantic analysis

These entity occurrences are encoded in a fairly common way, using <rs> elements which link to lists of elements in which the entities are described and linked to external authority records, and <date> elements in the case of chronological references. At a later project stage, we will explore the possibilities to derive other formats from this data which facilitate the extraction and processing of their semantic content, such as Geography Markup Language (GML)/Keyhole Markup Language (KML) for spatial data, or CIDOC-CRM for events. This paper will explore how our entity data, which is available in similar form in many other TEI

encoded editions, can be put to use in ways that go beyond the traditional uses described above, and which enter the realm of semantic analysis. Examples include:

- counting entities and calculating their relative frequency. We expect a high number and a concentration for pages where we can find short notations or lecture notes. Thus, we hope to be able to distinguish these parts from literary manuscripts;
- enriching personal data with birth and death dates from authority files and calculating differences in order to identify historical strata;
- identifying co-occurrences of persons and other entities and constructing networks in order to calculate graph theoretical measures;
- connecting places to routes, visualizing them on maps and calculating their distances using coordinates from external databases. Place entity references can occur in several different roles#: in this context, we must distinguish places visited by Fontane where he took notes, and distant places only mentioned by Fontane. It will be of interest to analyse the differences and similarities between these two geographic networks, particularly when a chronological dimension (i.e. the date of Fontane's visit, or the date of a historic event referred to by Fontane which took place at a mentioned site) is added;
- comparing Fontane's statements about entities, such as dates, locations, and names, with what we know about them today.

These data aggregations will be provided to the user as interactive graphics using D3.js or in the case of locations connected to a specified time or period, using the DARIAH GeoBrowser e4d#. Therefore we develop XSLT transformation scenarios, build with XQuerys within our exist-db (project portal), that delivers the needed JSON (D3.js) or KML (e4d[1]#) formats and transfer these data sets using appropriate interfaces.

Bibliography

- [1] James Cummings, "The William Godwin's Diaries Project: Customising and transforming TEI P5 XML for project

work", in: Jahrbuch für Computerphilologie 10 (2008), http://computerphilologie.de/jg08/cummings.pdf (April 29, 2009), last visited on March 27, 2013

- [2] Women Writers Project, "Women Writers Online", http://www.wwp.brown.edu/wwo/lab/speakers.html, last visited on March 27, 2013
- [3] Mike Bostock, "Force Directed Graph", http://bl.ocks.org/mbostock/4062045, last visited on March 27, 2013; based on data provided by Donald Knuth, "The Stanford GraphBase: A Platform for Combinatorial Computing", Reading 1993
- [4] Gabriele Radecke, "Theodor Fontanes Notizbücher. Überlegungen zu einer überlieferungsadäquaten Edition", in: Martin Schubert (Ed.), Materialität in der Editionswissenschaft, Berlin 2010 (= Beihefte zu editio; Bd. 32), pp. 95–106. – The Berlin State Library is the owner of the notebooks and an associated partner of the project.
- [5] Project website http://www.unigoettingen.de/de/303691.html and http://www.textgrid.de/community/fontane/
- [6] Gemeinsame Normdatei / Integrated Authority File of the German National Library, http://www.dnb.de/EN/Standardisierung/Normdaten/GND/gnd_node.html, last visited on March 27, 2013
- [7] Virtual International Authority File, http://viaf.org/, last visited on March 27, 2013
- [8] Humphrey Southall, "Defining and identifying the roles of geographic references within text: Examples from the Great Britain Historical GIS project", in: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1, pp. 69-78, doi:10.3115/1119394.1119405
- [9] europeana4D: exploring data in space and time, http://dev2.dariah.eu/e4d/, an example using the content from one single page can be found at http://goo.gl/TSNDf, last visited on March 27, 2013
- [10] EuropeanaConnect: "KML Specifications", http://tinyurl.com/e4d-kml, last visited June 27, 2013

Ontologies, data modelling, and TEI

Eide, Øyvind

Ontologies

In philosophy, *Ontology* denotes the study of being, with traces at least 2500 years back in history. In computer science, *ontologies*, uncapitalised and in the plural, has been a topic of study for some thirty years, initially connected to the artificial intelligence community. Computer science ontologies refer to shared conceptualisations expressed in formal languages (Gruber, 2009). They have not been of much importance in digital humanities before the last 10-15 years, but are now gaining momentum, connected to the development of the semantic web.

In the paper I will discuss ontologies in the context of the Text Encoding Intiative (TEI Consortium, 2012), based on the computer science tradition. However, even if computer science ontologies are different from philosophical Ontology, the two are not totally disconnected (Zúñiga, 2001) and some remarks will be made on links to philosophy as well. The focus will be on how meaning can be established in computer based modelling, in connection with the sources. Meaning can be based on the sources and the interpretation of them, but can also be established through the development of the ontologies themselves.

It is sometimes claimed that TEI expresses an inherent ontology, and in some sense it is true. TEI represents a shared conceptualisation of what exists in the domains relevant to text encoding. However, even if TEI can be expressed in formal models, it is questionable whether TEI can be seen as an ontology in the computer science sense. According to the classification in Guarino et al. (2009, 12–13), XML schemas are typically not expressive enough for the formality we need for ontologies. However, the level of language formality forms a continuum and it is difficult to draw a strict line where the criterion of formal starts. This continuum can be connected to different parts of the TEI. Some parts, such as the system of persons, places, and events, may be closer to an ontology than other, less formalised parts of the standard (Ore and Eide, 2009).

Two ways of modelling

There are no ontologies without models---an ontology, after all, represent a model of a world or of a certain corner of it. The discussion in the paper will focus on active engagement with models, that is, on how meaning is generated and anchored when ontologies and other models are developed and used. For TEI specifically, creating the standard was of course dependent on ontological consideration in the philosophical sense. Further, using it may also include similar ontological studies of the source material.

I will distinguish between two different, although overlapping, ways of modelling. First, one may use already existing models for data integration. An example of this is the task of integrating data from several different libraries and archives in order to create a common data warehouse in which the detailed classifications from each of the databases are preserved. In the process, one will want to use a common ontology for the cultural heritage sector, for instance, FRBRoo (FRBR, 2012). In the process, one must develop a thorough understanding of the sources, being they TEI encoded texts or in other forms, as well as of the target ontology---one will develop new knowledge.

The task is intellectually demanding and the people engaged in it will learn new things about the sources at hand. Still, the formal specification of the corner of the world they are working towards is already defined in the standard. Only in a limited number of cases will they have to develop extensions to the model. Once the job is done, making inferences in the ontology based data warehouse can be used to understand the sources and what they document even better. Yet, all the learning included, the process is still mostly restricted to the use of what is already there.

The second way of working with models is to create an ontology or another formal model through studying a domain of interest. In this case, a group of people will analyse what exists in the domain and how one can established classes which are related to each other. This may, for instance, be in order to understand works of fiction, as in the development of the OntoMedia ontology, [URL: http://www.contextus.net/ontomedia/model (checked 2013-03-30)] which is used to describe the semantic content of media expressions. It can also be based on long traditions of collection

management in analog as well as digital form, as in the development of CIDOC-CRM (CIDOC, 2011) in the museum community. Although one will often use data from existing information systems, the main goals of such studies are not mappings in themselves, but rather to understand and learn from previous modelling exercises in the area of interest.

The historical and current development of TEI can be seen in this context. The domain of TEI has no clear borders, but the focus is on text in arts and cultural history. In order to develop a model of this specific corner of the world, one had to analyse what exists and how the classes of things are related to each other. This is a process in which domain specialists and people trained in the creation of data models must work together, as the history of TEI is an example of.

When applying either of the two ways of modelling, knowledge is gained through the process as well as in the study and use of the end products; one can learn from modelling as well as from models, from the process of creating an ontology as well as from the use of already existing ones. It is a common experience that actively engaging with a model, being it in creating or in using it, gives a deeper understanding than just reading it. Reading the TEI guidelines is a good way of getting an overview of the standard, but it is hard to understand it at a deeper level without using it in practical work, and it is quite clear that among the best TEI experts are those who have taken part in creating the standard.

There is no clear line between the two ways of modelling, and they often use similar methods in practice. They both have products as the end goal, and new knowledge is created in the process. Some of this new knowledge is expressed in the end products. For example, working to understand better what is important for a concept such as ``person'' in the domain used will results in new knowledge. This knowledge will be shared by the parties involved and may be expressed in the end product. However, there is a stronger pressure towards expressing clearly such new knowledge when a data standard is created than when a mapping is created.

Interconnections

An ontology may or may not include contradictory facts, and may contain them at different levels. How this can be related to different interpretations

of the source material will be discussed in the paper and differences between TEI and ontologies such as CIDOC-CRM will be pointed out.
While an ontology is a model of the world, a specific mapping to an ontology will be based on sources. Ways of linking ontologies to their sources in order to ensure scholarly reproducibility will be presented in the light of co-reference and of links between text encoding and ontologies in general. As a case study, this will be done through a study of ways of linking TEI to CIDOC-CRM. While the two standard will continue to develop, and in some areas, such as person, place, event, and possibly object, they may grow closer, they will still continue to be two separate standards, different in scope as well as in the ways in which they are formalised.
The paper will investigate into various ways of interconnecting the two as part of modelling work, and develop a draft categorisation of the most common types. I will be looking forward to receiving feed-back from a qualified audience on the draft system in order to develop it further.

Bibliography

- CIDOC (2011). *Definition of the CIDOC Conceptual Reference Model.* [Heraklion]: CIDOC. Produced by the ICOM/CIDOC Documentation Stan- dards Group, continued by the CIDOC CRM Special Interest Group. Ver- sion 5.0.4, December 2011.
- FRBR (2012). *Object-oriented definition and mapping to FRBR(ER) (Version 1.0.2).* [Heraklion]: International Working Group on FRBR and CIDOC CRM Harmonisation. “The FRBRoo Model”.
- Gruber, T. (2009). Ontology. In L. Liu and M. T. Özsu (Eds.), *Encyclopedia of Database Systems,* pp. 1963–1965. [S.n.]: Springer US.
- Guarino, N., D. Oberle, and S. Staab (2009). What Is an Ontology? In S. Staab and R. Studer (Eds.), *Handbook on ontologies,* pp. 1–17. Berlin: Springer. 2nd ed.
- Ore, C.-E. S. and Ø. Eide (2009). TEI and cultural heritage ontologies: Exchange of information? *Literary & Linguistic Computing* 24(2), 161– 172.

- TEI Consortium (2012). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. [2.1.0]. [June 17 2012]*. [S.n.]: TEI Consortium.
- Zúñiga, G. L. (2001). Ontology: its transformation from philosophy to information systems. In N. Guarino, B. Smith, and C. Welty (Eds.), *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001,* pp. 187–197. Ogunquit, Maine, USA: ACM.

TEI and the description of the Sinai Palimpsests

Emery, Doug; Porter, Dot

The library of the Monastery of St. Catherine's in the Sinai Desert is well known as the source of Codex Sinaiticus and the home of the palimpsest Syriac Sinaiticus, both of which date to the 4th Century C.E. It also preserves a collection of 120 known palimpsests in Greek, Syriac, Georgian, Armenian, Arabic, and several other languages. Few of these had been studied extensively. The same team of technical experts, engineers and scientists responsible for imaging the Archimedes Palimpsest, the Galen Syriac Palimpsest, the Waldseemüller 1507 World Map, and David Livingstone's 1871 Field Diary are now producing enhanced images of the original undertext in the Monastery's palimpsests. After a 2009 technical survey by the team, in 2011 a five-year project began to image and survey the palimpsests at the monastery in a collaboration of St. Catherine's Monastery and the Early Manuscripts Electronic Library. This latest project builds on the team's previous spectral imaging work, which pioneered the use of spectral imaging techniques in several modalities to collect manuscript data and produce processed images to enhance the visibility of the erased undertexts.

The project is also responsible for documenting the physical condition of the manuscripts, each palimpsest folio, and identifying the texts inscribed in each undertext layer. To encode the very complex descriptions of the manuscripts and their undertext layers, the project will need to employ the TEI.

This paper will discuss the Sinai Palimpsest Project's use of the TEI to describe the palimpsests, building on the methods developed in previous projects including the Archimedes Palimpsest, Livingstone Diary, and the Walters Art Museum's series of NEH-funded manuscript preservation and access projects.

It will also provide a survey of methods employed and challenges encountered. Most importantly, it will elicit advice and suggestions for future TEI use, and identify areas where the TEI may need to be modified to aid in complex palimpsest descriptions.

The palimpsests at St. Catherine's have varied and complex structures. Some folios have been reused more than once, so that in the collection there are several double-palimpsests, and even some triple-palimpsests with multiple layers of scraped or washed off text. The orientations of undertext to overtext vary from manuscript to manuscript, and even within a single manuscript. Some leaves were created by stitching together portions of reused folios, so that some present-day leaves are literal palimpsest patchworks. These conditions present challenges not only for scholars reading the undertexts, but also for their system presentation by computer applications.

The Sinai Palimpsests Project employs a complex model for describing palimpsest structure. Each manuscript has a number of palimpsest folios. Each folio may have one or more undertext layers. Participating scholars are assigned sets of undertext layers from a manuscript, grouped by language and script, based on each scholar’s area of expertise. Some manuscripts have undertext layers in several languages and scripts, and, thus, have several under text layer groupings. The scholar examines each folio undertext layer in the assigned grouping and links the undertext layer to an "undertext object". An undertext object is a collection of folio undertext layers that have the same textual content and are written in the same hand. The requirement for an undertext object is rather strict. For

example, folio undertext layers written in the same hand, but belonging to two separate New Testament books would be assigned to two undertext objects. By this method each manuscript is divided by language and script, and then digitally sorted into undertext layers that likely belonged together in the same 'original' manuscript.

In a second level of analysis, scholars will examine undertext objects to determine which ones originally belonged together and link them together. Linked undertext objects may be from the same present-day manuscript or, as will often be the case, from separate present-day manuscripts. An example of this is ongoing studies of the Syriac Galen Palimpsest, which appears to have leaves scattered about the globe. These leaves are just now being tracked down by scholars. If the assumption is correct – that many of these palimpsests were generated at Sinai – it is likely that leaves from a number of manuscripts were reused and spread across two or more later manuscripts.

The TEI structure used to describe the palimpsests must express this complexity. The resulting TEI encoding will identify each undertext work, and, where possible, describe the reconstructed undertext manuscripts. Doing so will require reconstructing undertext folios from palimpsested pieces that often span more than one present-day folio. The TEI encoding will integrate the project's manuscript images; and undertext folio descriptions should map to that image data as much as possible. One goal of the project is to provide TEI manuscript descriptions that will allow applications to display images of folios in their current overtext or the reconstructed undertext form and order. Using encoded information about a palimpsest, such a tool should be able to select an image of a folio, rotate it to display an undertext layer the right way up, and if need be join that image with another image or images to present a reconstructed view of the original undertext folio. In the case of patchwork folios, the tool should be able to select the portion of an image corresponding to an undertext layer. The markup that supports these functions should provide the following.

- A list of the overtext folios
- – A description of the undertext content
  – The undertext orientation, relative to the overtext

  - The layout of the undertext (columns, number of lines)
  - The portion of the undertext folio preserved (bottom half, top half, lower left quarter, etc.)
- For "patchwork" folios, a method for designating a region of a folio as an undertext layer and linking that undertext layer to a region of an image
- A method for linking several undertext layers together as parts of a single undertext folio
- A method for collecting several undertext folios together as part of a reconstructed "undertext manuscript", which will have its own manuscript description

The complexity of the problem raises the question of whether a single TEI file can adequately and fully describe a manuscript and its undertexts, or whether this information can even be encoded in the TEI alone. One approach would be to create separate TEI files for each present-day manuscript, and then one for each reconstructed undertext manuscript. This approach solves the problem of dealing with undertext manuscripts that span several modern ones, but it necessitates markup that spans files to express relationships between over- and undertext folios. An alternate method would create a single TEI file for each current manuscript, including over- and undertexts, by assigning each reconstructed manuscript to its own TEI msPart. If the use of the TEI proves unwieldy for some features, a custom standoff markup, linked to the TEI may be used to encode complex overtext and undertext relationships. This paper will give examples of each method.

The volume, complexity and variety of the Sinai palimpsests provide a unique opportunity to explore the use of the TEI for palimpsest descriptions in support of global virtual scholarly studies. TEI can serve as a key tool in this and other scholarly studies of complex texts with well researched and documented application of the opportunities and utility of the TEI to support scientific, technical, and scholarly applications to digital humanities.

Bibliography

- Bockrath, Diane E., Christopher Case, Elizabeth Fetters, and Heidi Herr. "Parchment to Pixel: The Walters Islamic Manuscript Digital Project." *Art Documentation*, 29, no. 2 (2010): 14-20.
- Emery, Doug, Alexander Lee, Michael B. Toth, "The Palimpsest Data Set", *The Archimedes Palimpsest, I. Catalogue and Commentary* (Cambridge: Cambridge University Press), pp. 222-239.
- Emery, D, F.G. France, and M.B. Toth, “Management of Spectral Imaging Archives for Scientific Preservation Studies”, Archiving 2009, Society for Imaging Science and Technology, May 4-7 (2009), 137-141
- Emery, D., M. B. Toth, and W. Noel, 'The convergence of information technology and data management for digital imaging in museums', *Museum Management and Curatorship*, 24: 4, 337 — 356 (2009)
- Porter, Dot, “Facsimile: A Good Start,” TEI Member’s Meeting, King’s College London, November 2008.
- TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.3.0. 17 Jan 2013. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (30 March 2013)

From TUSTEP to TEI in Baby Steps

Fankhauser, Peter; Pfefferkorn, Oliver; Witt, Andreas

The "Mannheimer Korpus Historischer Zeitschriften und Zeitungen"[1] (MKHZ) aims at documenting German news language in the 18th and 19th century. The corpus is available both, as high resolution jpeg files

[1] Mannheim Corpus of Historical Magazines and Newspapers, PID: http://hdl.handle.net/10932/00-01B8-AE41-41A4-DC01-5

and as TUSTEP transcriptions that have been acquired in a double keying procedure. The current version of the corpus comprises 21 magazines with 652 individual volumes with over 4.1 Mio word tokens on 4678 pages.

In this paper we briefly describe the original TUSTEP markup available for MKHZ and introduce an iterative and staged pipeline for transforming TUSTEP markup to TEI. The pipeline is set up in three main stages: (1) Syntactic transformation of TUSTEP to well-formed TUSTEP XML, (2) transformation of TUSTEP XML to generic TEI, (3) refinement of generic TEI with magazine specific logical structure.

The corpus has been transcribed using TUSTEP conventions[2]. TUSTEP [2] is a framework for compiling critical editions. It predates XML, and parallels SGML, as the predecessor of XML. The main unit of a TUSTEP transcription is a numbered line. For the MKHZ corpus the TUSTEP markup represents layout structure (lines, columns and pages), logical structure (paragraphs with alignment information, tables, figures, running headers, and footnotes), typographic information (font family, style, and size), and special symbols (mostly glyphs), numbers, etc.

The layout structure is fairly complex. In particular advertising sections make heavy use of multiple, possibly nested columns, which do not necessarily range over an entire page. In contrast, the marked up logical structure is fairly simple. There exists no explicit distinction between headings and ordinary paragraphs, though heuristic rules based on style information, such as text-alignment or typography can be used to differentiate between these elements. Moreover, individual articles and their sections are not marked up explicitly. Altogether the TUSTEP markup of MKHZ focusses on layout structure and typographic annotation, which is translated to TEI in three main stages:

(1) In the first stage the TUSTEP markup is transformed to well-formed XML, which reflects the original markup as closely as possible[3], without losing any markup and content or introducing spurious markup. This comprises two main challenges: Firstly, TUSTEP employs a significantly

[2] A small portion of the corpus has been transcribed independently (together with other resources) in the GerManC project [1]. In this project however, the original transcription did not use TUSTEP and was enriched directly with TEI markup manually

[3] This first stage is readily comparable to the approach described in [3]. However, other than [3], we aim at a lossless transformation of TUSTEP to XML.

more diverse markup syntax than XML, and secondly, it interleaves layout structure with logical structure and makes liberal use of tag omission.

To capture TUSTEP's diverse syntax, we extract and iteratively refine markup patterns and specify their translation to XML markup. To resolve conflicts between layout structure and logical structure, we break up logical elements, such as paragraphs and tables, and insert continuation milestones to link the broken up elements with each other.

Technically, this stage is implemented in Perl[4], as a pipeline of custom event-based parsers, one for producing basic well-formed XML, and one for transforming tabulated tables into tables consisting of rows and cells. Where tag omissions or wrong markup cannot be resolved automatically, the original TUSTEP markup is modified and documented in the form of a diff list. From the resulting XML we generate and manually refine an XML-Schema to validate the output and guide the transformation in Stage 2.

(2) The adhoc XML vocabulary resulting from Stage 1 is rather complex, comprising about 50 elements. This complexity is deliberate, because it allows for a fine-grained check of markup balance based on XML's well-formedness criterion. In Stage 2 this complexity is reduced by mapping the vocabulary along TEI guidelines [6]. Typographic markup is transformed to highlight elements with style attributes, structural markup is unified to paragraphs with appropriate style and type attributes, and all other elements are mapped to appropriate TEI elements. Moreover, the continuation milestones introduced in Stage 1 are used to link separated logical elements by means of so called virtual joins along the guidelines in [6, Section 20.3].

Technically, this stage is implemented as a pipeline of XSLT scripts, one for mapping to TEI, followed by one for inserting virtual joins. The result of this stage is TEI compliant markup, which still represents the original markup without information loss, but largely differentiates by means of

[4] [4] describes an iteratively refined custom transformation from TUSTEP to TEI-SGML by means of TUSTEP's script language TUSCRIPT, conceptually very similar to the transformation pipeline presented in this paper. Aiming to use standard (and familiar) technology, we have chosen to split the transformation to TEI into a pipeline of Perl feeding XSLT. We have also investigated the use of TXSTEP [5], which aims at providing TUSCRIPT and TUSCRIPT-modules in an XML-Syntax. However, the available modules did not cover the needs of the transformation at hand.

attributes rather than elements, resulting in a significantly less complex schema.
(3) The final stage aims at explicating hidden logical structure, in particular identifying independent articles within an issue and capturing meta information such as issued date. This requires heuristic rules specific for the 21 individual magazines. The rules use local context information such as (usually centered) headings and typographic patterns to group sequences of paragraphs into articles. This final transformation is carried out by means of iteratively refined custom XSLT scripts and manual annotation.
In summary, the presented pipeline aims at managing the complexity of transformation by dividing it into several stages, which can be individually refined and validated. Each stage simplifies and unifies the markup and underlying model, making the subsequent stage more tractable. The modular structure of the pipeline also facilitates its adaptation to other TUSTEP sources. However, especially the mapping from TUSTEP XML resulting from stage 1 to TEI probably requires adaptations to the particular TUSTEP vocabulary at hand.
The resulting TEI representation is used as a pivot model for generating a visualization in xhtml + css, which closely reflects the original layout structure, for extracting meta data as a basis for archiving the corpus in the IDS Repository [7], and for generating a representation in the IDS Text Model [8] for import into the Corpus Search and Analysis System COSMAS II [9].

Bibliography

- [1] Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett: Annotating a historical corpus of German: A case study. Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards", Valletta, Malta, 18 May 2010. 64-68.
- [2] Universität Tübingen; Zentrum für Datenverarbeitung. TUSTEP 2012: Handbuch und Referenz (electronic Version, in German). Available at: http://www.tustep.uni-tuebingen.de/
- [3] René Witte, Thomas Kappler, Ralf Krestel, and Peter C. Lockemann: Integrating Wiki Systems, Natural Language

Processing, and Semantic Technologies for Cultural Heritage Data Management, In: Language Technology for Cultural Heritage, pp.213-230, Springer, 2011.

- [4] Ruth Christmann: Books into Bytes: Jacob and Wilhelm Grimm's Deutsches Wörterbuch on CD-ROM and on the Internet. http://germazope.uni-trier.de/Projekte/DWB/bibliographie/books.htm (accessed March 23, 2013)
- [5] Wilhelm Ott, Tobias Ott, Oliver Gasperlin. TXSTEP – an integrated XML-based scripting language for scholarly text data processing. Digital Humanities 2012.
- [6] TEI Consortium, eds.: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.3.0. Last updated on 17th January 2013. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (accessed March 23, 2013).
- [7] Peter M. Fischer, Andreas Witt. Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache. In: Proceedings of the LREC 2012 Workshop 'Best Practices for Speech Corpora in Linguistic Research', Istanbul, May 21, 2012 (pp. 47-50). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/workshops/03.Speech%20Corpora%20Proceedings.pdf
- [8] Harald Lüngen, C.M. Sperberg-McQueen: A TEI P5 Document Grammar for the IDS Text Model In: Journal of the Text Encoding Initiative (2012), H. 3. http://jtei.revues.org/508
- [9] Franck Bodmer: COSMAS II. Recherchieren in den Korpora des IDS. In: Sprachreport 3/2005. S. 2-5 - Mannheim: 2005.

How TEI is Taught: a Survey of Digital Editing Pedagogy

Gavin, Michael Andrew; Mann, Rachel Scott

One of the remarkable shifts in the field of humanities computing and the digital humanities has been its emergence in recent years as a topic of instruction across universities in Europe and North America. From a cluster of specialized research techniques, humanities computing is increasingly encountered in the classroom as a subject of scholarly discussion in its own right. In humanities education, the boundary between "content" and "skills" has long been blurry and contested, and the rapid increase in courses devoted to digital humanities is testing that boundary in new and exciting ways. TEI holds a significant place within this larger picture. In Lisa Spiro's 2011 survey of 134 digital humanities syllabi, XML and TEI were, by an overwhelming margin, the most frequently taught technologies (Spiro). In workshops, seminars, general courses in the digital humanities, and specialized courses on digital editing with "angle bracket technologies," students encounter TEI both as a set of skills to master and as a topic among other topics within a disciplinary field of knowledge. TEI is taught in diverse formats to diverse audiences. In this talk, we will present an overview of TEI pedagogical documents (course syllabi, workshop and seminar descriptions, and instructional materials) as well as the results of our ongoing survey of TEI instructors. Our purpose will not be prescriptive nor predictive; that is, we will not outline a program for how TEI should be taught nor provide directions for the future. Instead, our purpose is simply to provide a picture, with as much detail as possible, of the state of TEI in 2013 from the perspective of the classroom.

TEI on the Syllabus

In our preliminary survey of TEI instructors, 52% of respondents reported teaching TEI in college courses devoted in part or in whole to digital editing. Our presentation will focus on syllabi and course descriptions that include TEI in order to see how TEI is practiced and imagined

across disciplines and departments. The syllabi range from English, history, digital humanities, information technology, and library and information science. XML and TEI often feature prominently in digital humanities courses, where they tend to be studied alongside media theory and computational analysis. DH instructors often lead one- or two-day sessions in the middle of the semester on TEI, which is then used as the format for class projects. In this context, TEI work is often described as a "practical" use of "tools" within the DH curriculum (for example, Galey). Through the construction of research protocols, attention to cultural histories, and "major epistemological, methodological, technological, and institutional challenges" (Presner), students are exposed to TEI as a tool with which to understand, know, and explore the products of culture. In addition to providing a framework for undergraduate research, XML is increasingly presented to graduate students as a part of their introduction to digital work, sometimes on the belief that it is less likely than other digital formats to become obsolete (Reid). In the field of Library Science, TEI is written into course descriptions and syllabi as having both practical and theoretical aspects worth considering, yet hands-on practice is, by and large, at the fore. Information science courses, such as "Advanced XML: Electronic Publishing Standards and Systems" (Walsh) and "Information Modeling in XML" (Binkley), tackle advanced technical skills like XSLT and linked data. On the other hand, courses like "Seminar in Historical Editing in the Electronic Era," taught in a history department, foreground the editorial questions and problems sparked by digital remediation (Katz; see also Rehbein and Fritze). Our discussion will provide an overview of our syllabus collection as a whole and analyze pertinent examples of general trends. Our emphasis will be on the most recent courses, and we expect our body of data will change significantly when Fall 2013 courses are announced.

The Workshop as a Genre of TEI Instruction

An important genre of TEI instruction continues to be the workshop or seminar, typically lasting from 1 to 5 working days. Workshop series hosted by Oxford and Brown have reached a wide community of students. Oxford's Summer 2012 TEI workshop offerings ranged from introductory surveys in which basic mark-up and TEI Guidelines and approaches to

publishing TEI texts were addressed to more advanced workshops in which students learned how to transform their TEI XML texts into texts other than HTML. With the help of NEH funding, Brown offered a series of TEI workshops to 11 North American universities from January 2007 to June 2009. Project director Julia Flanders describes their goal to teach "text encoding in a way that emphasizes its theoretical and methodological significance." Elena Pierazzo explains that workshops taught at King's College London are founded on the belief that "students want to do better or new research." Teaching strategies include the incorporation of attendee-brought material, exercises relevant to that material and the introduction of resources that will enable attendees to become self-sufficient after completion of the course. Workshops hosted by the Japanese Association for Digital Humanities and the University of Paderborn's Edirom Summer School 2012 foregrounded the acquisition of markup skills and the "independent handling" of TEI guidelines. Across students' variety of interests and motivations, the primary challenge for workshop-based instruction is, in James Cumming's words, to "produce a consistent pedagogical basis while retaining their unique character and experiences."

2013 Survey of TEI Instructors

Our discussion will also provide an overview of responses to our "Teaching TEI" survey, a preliminary version of which was distributed this spring, receiving more than 30 responses from TEI instructors in Europe, North America, and Japan. This survey will continue to be available over the summer and will be updated next fall.
In the survey we ask:

- 1) In what country and language do you primarily teach?
- 2) In what language do you primarily teach? 3)
- What is your position within your institution?
- 4) What is your home department or administrative unit?
- 5) What year did you first teach digital editing, XML, or TEI?
- 6) How frequently do you teach digital editing, XML, or TEI?
- 7) In what format do you teach digital editing, XML, or TEI?

- 8) When you teach digital editing, XML, or TEI, who is your primary audience?
- 9) Were you financially compensated for your extracurricular teaching?
- 10) Have you ever charged a fee to participate in a workshop?
- 11) Do you create your own course materials? What textbook or other resources do you use?

We also invited respondents to list courses and workshops taught and to describe their experience in their own words, which has allowed us to gather significant testimony from instructors new to the field.

Like the TEI community, our respondents are diverse, whether by country, language, or discipline. Our talk will provide a detailed breakdown of responses. Perhaps the most intriguing line of distinction we have found so far is years of experience. For many, teaching TEI is a new addition to their scholarly work. When asked what year they began teaching, the single most frequently reported year is 2012. Of our respondents to date, the median experience is 6 years, with a fairly even split of about 30% each between those who have taught TEI for more than eight years and those who began only since 2011. These two groups are very different. Within our set, new teachers are far more likely to teach TEI as part of a college course curriculum and much less likely to teach workshops. Their target audiences are much less likely to include professors and university staff and are more likely to be limited to undergraduates and graduate students within their respective disciplines. New teachers among our respondents are much more likely to be faculty in a literature or history department and much less likely to be library or IT professionals.

These results are consistent with our general picture: TEI is increasingly being taught and understood as a component of the general humanities curriculum. This change marks TEI's pedagogical success and its growth in size and scope. This also means, however, that the audience of TEI pedagogy is increasingly an undergraduate audience, and that research projects completed in TEI will often take shape in the classroom. Meeting the needs of this growing audience and its research demands is one of the most important challenges facing the TEI community today.

Bibliography

- Binkley, P. (2012). LIS 598 - Information Modeling in XML - University of Alberta. http://www.slis.ualberta.ca/en/Courses/GraduateCourses/LIS598InfoModelXMLOutline.aspx
- Cummings, J. (2012). Teaching the TEI Panelhttp://blogs.oucs.ox.ac.uk/jamesc/2012/11/22/teaching-the-tei-panel/
- Cummings, J., Baalen, R., and Berglund-Prytz, Y. (2012). An Introduction to XML and the Text Encoding Initiative.http://digital.humanities.ox.ac.uk/dhoxss/2012/workshops.html
- Flanders, J. (2009). Final Report: Seminars in Humanities Text Encoding with TEI. http://www.wwp.brown.edu/research/publications/reports/neh_2007/seminars_report.html
- Galey, A. (2009, Winter). FIS 2331H: Introduction to Digital Humanities. University of Toronto.
- Hawkins, K. (2012, November 17). Creating Digital Editions: An Introduction to the Text Encoding Initiative (TEI). http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei
- Katz, E. (2012, Fall). Historical Editing in the Digital Era. New York University.
- Mahony, S., and Pierazzo, E. (2012). Teaching Skills or Teaching Methodology? In B. D. Hirsch (Ed.), Digital Humanities Pedagogy. Open Book Publishers.
- Pierazzo, E. (2011, September 21). Digital Editing. Elena Pierazzo's Blog. http://epierazzo.blogspot.com/2011/09/digital-editing.html
- Pierazzo, E., Burghart, M., and Cummings, J. (2012). Teaching the TEI: from training to academic curricula. TEI Conference, College Station, TX. http://idhmc.tamu.edu/teiconference/program/papers/#teach
- Presner, T. (2012, Winter). Introduction to the Digital Humanities. http://introdh.blogspot.com/p/syllabus.html
- Rehbein, M., and Fritze, C. (2012). Hands-On Teaching Digital Humanities: A Didactic Analysis of a Summer School Course

on Digital Editing. In B. D. Hirsch (Ed.), Digital Humanities Pedagogy. Open Book Publishers.

- Reid, A. (2012) Graduate Education and the Ethics of the Digital Humanities. In M. Gold, (Ed.), Debates in the Digital Humanities. University of Minnesota Press.
- Reid, A. (2012) Graduate Education and the Ethics of the Digital Humanities. In M. Gold, (Ed.), Debates in the Digital Humanities. University of Minnesota Press.

TEI metadata as source to Europeana Regia – practical example and future challenges

Gehrke, Stefanie

Europeana Regia (2010-2012) was a project co-funded by the European Commission in the context of the Europeana project. Focusing on the incorporation of digitised manuscripts from the Middle Ages to the Renaissance into Europeana, it was the aim to make the manuscripts of the Carolingian period (Bibliotheca Carolina), the library at the Louvre in the time of Charles V and Charles VI (Library of Charles V and Family) and the library of the Aragonese Kings of Naples virtually accessible.
The source metadata at the participating institutions was available in multiple formats (e.g. MARC21, EAD and TEI) and in different levels of detail, while the Europeana format in the beginning of the project was ESE v3.2. A lot more was needed, than just producing valid records: In order to compile the digital facsimiles via Europeana into unique virtual collections, for Europeana Regia manuscripts a certain specification of the ESE (Europeana Semantic Elements) metadata was agreed on. Considering that each medieval manuscript is a unique piece of work and also having in mind, that the individual or institution responsible for the encoding of the metadata might have their own approach to the matter,

it became obvious, that the task not only needed standards but also some way to check if any set of metadata would fulfil these standards to assure high quality within the project.

In order to identify the standard, a complete crosswalk of the necessary information for display in Europeana encoded in ESE v3.4 for the different input formats was compiled. While some partners already had academic metadata of the manuscripts, e.g.in TEI, others still had to choose a metadata format to encode their lists and free text descriptions in. Furthermore one has to keep in mind, that especially for high level encoding like TEI, there are often multiple ways to express the same relation or content (<head><title> vs. <summary> ; <rs type=”person”> vs. <persName>). So in the end, apart from long lists, the true crosswalk within all the encodings used in the project was represented by a single reference transformation to ESE combining all different input format modules with a single Europeana output module. It served the purpose of quality insurance tool prior to and after ingestion as well. For TEI this meant, that the reference transformation would need a certain subset of TEI as suitable input for medieval manuscripts. Finally ENRICH compliant TEI was used with a few additions. For institutions that already have a lot of TEI metadata that is not ENRICH compliant a path to Europeana can be implemented that first creates a reduced export metadata set which then is transformed to ESE.

The XSLT code of the reference transformation is clearly structured, well commented and expandable to accommodate for further input formats like METS/MODS. The paper will show the key elements of this export metadata format and how it maps to the ESE fields and the final display in Europeana. From other examples it becomes obvious, that encoding should always take the most advantage from the encoding format used, as tagged metadata is much simpler read by machines than text format conventions in the content entries. The standards of the Europeana Regia project actually exceeded the necessities of the ESE format by the used of identifiers, that are not properly used in the ESE context but already point to the semantic future of Europeana with EDM.

But the Europeana Regia project was more than just an effort in digitisation, creation of metadata and ingestion into Europeana. On

the Europeana Regia portal (www.europeanaregia.eu) the partners also provided translations from the original metadata language to all languages of the participating institutions. This multilingual metadata content is still a treasure that needs to be incorporated into Europeana as well as the use of identifiers and authoritative data.

This leads to the future of Europeana as a retrieval tool for the semantic representation also of the fully linked medieval manuscript metadata with EDM (Europeana Data Model). While EDM especially for manuscripts is still a work in progress (DM2E), a lot can be learned from and already be done based on the Europeana Regia work. The author will show how the reference transformation was changed to produce valid EDM from TEI, MARCXML and EAD. For TEI the advantages and the caveats will be presented, when trying to make full use of the semantic EDM – RDF possibilities, based on academic metadata for medieval manuscripts encoded with TEI. A description of the reference transformation to EDM is given based on the XSLT code and metadata examples from the project. While the representation of manuscript metadata in sematic ontology gains momentum the author hopes to provide some suggestions on future TEI use in that field.

Bibliography

- Europeana
- Europeana Regia
- Europeana Semantic Elements specifications v3.4.1
- Definition of the Europeana Data Model elements
- ENRICH Project

Documenter des “attentes applicatives” (processing expectations)

Glorieux, Frédéric; Jolivet, Vincent

Il est fréquent d’entendre dans la communauté TEI que l’encodage ne doit pas se soucier des traitements (*processing*). C’est une évidence : l’encodage ne doit pas dépendre de contraintes applicatives, mais bien de l’analyse du texte encodé et de ses composants. Un tel principe a certainement préservé TEI de plusieurs modes technologiques passagères. Il a aussi contribué à un accroissement important du nombre d’éléments (plus de 600 aujourd’hui). Pour autant, nous pensons que ce développement gêne maintenant le déploiement de TEI et qu’il se paie en complexité d’apprentissage, et d’implémentation. La documentation d’“attentes applicatives” (processing expectations) pour chaque élément nous semblerait utile à l’évaluation des définitions proposées par la TEI et favorable à la convergence des pratiques d’encodage.

Puisque TEI sert à encoder des textes, l’attente applicative la plus commune est sans aucun doute la lecture. Il est essentiel de pouvoir distribuer les textes encodés dans des formats utiles au lecteur : à l’écran (HTML, epub, etc.) ou imprimables (LaTeX, ODT, etc.). Le confort de (re)lecture est aussi déterminant dans le processus de correction des corpus textuels encodés. Pour cela, nous utilisons tous les transformations maintenues par le consortium et l’outil de conversion OxGarage. Mais pour atteindre la qualité éditoriale attendue dans le milieu universitaire, nous sommes presque toujours contraints de les personnaliser, y compris pour des composants textuels aussi communs que les notes (d’auteur, d’éditeur, d’apparat critique) ou un index. La qualité des outils maintenus par le consortium n’est absolument pas en cause, mais plutôt la permissivité revendiquée de la TEI : « For the Guidelines to have wide acceptability, it was important to ensure that : (…) multiple parallel encodings of the same feature should be possible. » Un tel principe garantit l’adaptabilité de la TEI à presque tout type de sources et de questionnements scientifiques et explique sûrement son succès académique. Cette permissivité est une qualité incontestable de la TEI.

Pour autant, en autorisant des solutions d'encodage concurrentes pour un même besoin, elle complique les traitements, l'échange et les exploitations scientifiques : comment générer par exemple des index croisés d'entités nommées sur des fichiers TEI aux encodages hétérogènes ? Même les composants textuels les plus fréquents tel que le rendu typographique sont affectés (il n'existe pas de valeurs normalisées pour l'attribut **`@rend`**). C'est une fausse liberté qui est donnée ici, car ces composants les plus fréquents sont aussi les mieux connus et définis. Une attente applicative aussi élémentaire que celle de l'affichage pousse ainsi à mieux préciser le modèle textuel.

Notre proposition consisterait à ajouter pour chaque élément des *guidelines*, en plus de la définition et des exemples, une section "attentes applicatives" (*processing expectations*). Ces dernières concernent aussi bien l'affichage, que l'échange (`<teiHeader>`), ou l'exploitation scientifique (entités nommées, linguistique, etc.). On préciserait par exemple qu'un `<author>` dans un `<titleStmt>` désigne l'auteur principal du texte, contrairement à un `<author>` dans le `<sourceDesc>` ; qu'un élément `<persName>` dans un `<body>` peut alimenter un index des personnes citées, par regroupement de clés (**`@key`**, **`@ref`** ?). Des équivalents vers différents formats d'import et d'export, comme les traitements de textes, Dublin Core, HTML5, ePub, ou LaTeX pourraient illustrer ces "attentes applicatives" et préciser utilement la sémantique TEI par comparaison à d'autres formats. En développant différents outils (odt2tei, teipub, lateix), nous avons ainsi été contraints de faire des choix qui ne sont pas purement techniques, mais sémantiques, et pour lesquels nous aurions apprécié des orientations plus explicites. Une telle information existe, mais elle est éparpillée dans les chapitres de prose, sur la liste de diffusion ou exprimée implicitement dans les transformations maintenues par le consortium. Il serait précieux de la concentrer sur les pages de la documentation que nous consultons le plus quotidiennement (celles des éléments).

Sans être prescriptives, ces "attentes applicatives" favoriseraient la convergence des pratiques d'encodage des textes et, parce qu'il est difficile de reprendre des fichiers à l'encodage hétérogène et mal documenté, en amélioreraient la pérennité. Elles seront aussi de précieuses

indications pour les développeurs de logiciels, afin d'implémenter les fonctionnalités souhaitées par la communauté.

The Lifecycle of the DTA Base Format (DTABf)

Haaf, Susanne; Geyken, Alexander

Introduction

This paper describes a strict subset of TEI P5, the DTA 'base format' (henceforth DTABf, [DTABf]), which provides tagging solutions for the richness of encoding non-controversial structural aspects of texts while allowing only minimal semantic interpretation. While the focus of Geyken et al. (2012) was put on a comparison of DTABf with other commonly used XML/TEI-schemas such as TEI Tite, TEI in Libraries or TEI-Analytics, this article places particular emphasis on the lifecycle of DTABf.

The DTABf has been created in order to provide homogeneous text annotation throughout the corpora of the Deutsches Textarchiv (German Text Archive, henceforth: DTA, [DTA]). The goal of the DTA project is to create a large corpus of historical New High German texts (1600–1900) that is balanced with respect to the date of their origin, text type and thematic scope and thus is supposed constitutes the basis of a reference corpus for the development of the New High German language. As of June 2013, the DTA corpora contain 1363 works. The text basis is continuously extended either by texts digitized by the DTA or originating from other project contexts ([DTAE]).

The DTABf had been created by applying encoding recommendations formulated by the DTA to the texts digitized during the first project phase (2007–2010, 653 texts). On the basis of the resulting annotations it underwent a thorough revision where the handling of structural

phenomena was reconsidered and consistent solutions were determined. As a result of these efforts the DTABf now consists of three components:

- an ODD file specifying constraints on TEI elements, attributes and values thus reducing the flexibility of the TEI P5 tag set while still providing a fully TEI P5 conformant format ([DTABf ODD]);
- an RNG schema generated from that ODD ([DTABf RNG]);
- a comprehensive documentation explaining the handling of common structuring necessities as well as of special cases, and illustrating each phenomenon with examples from the corpus ([DTABf]).

The DTABf currently (June 2013) consists of 77 <teiHeader> elements and 50 <text> elements together with a limited selection of attributes and values (where feasible). The DTABf-header elements specify bibliographic information about the physical and the electronic source, the text classification and legal status of the document as well as information about the encoding. The DTABf-text elements include annotations of formal text structures (e.g. page breaks; lists; figures; physical layout information, such as form work, highlighting, etc.) as well as semantic text structures (heads; proper names; text types, such as poem, chapter, letter, index, note, etc.). Furthermore, the DTABf allows for documented editorial interventions (e.g. correction of printing errors or editorial comments). Linguistic information is not encoded inline since it is gained automatically and applied to the DTA texts via a standoff markup (Jurish 2010). The search engine of the DTA supports linguistic queries and allows filtering by DTABf elements and attributes ([DDC]).

The DTABf's Life Cycle

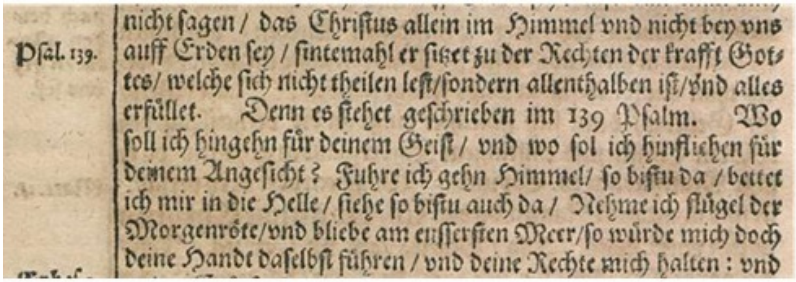
Despite the large and heterogeneous data basis upon which the DTABf has been built in the past years new structural phenomena may appear with new texts mainly because of individual printing habits in older historical works. In addition the markup of texts encoded by external projects may differ from the DTABf either formally or semantically. Therefore, the DTABf continually comes under scrutiny, the challenges being first to decide whether adaptations to the format are unavoidable in order to meet new requirements, and second, to ensure that such adaptations do not lead

to inconsistencies of the structural markup within the corpus, the latter being a necessary prerequisite for interoperability of the corpus resources. In the next section we illustrate these cases.

New Phenomena in the Scope of DTABf

New phenomena that are in the scope of DTABf fall into two classes: either a tagging solution relying on DTABf elements, attributes and values can be found for the structural phenomena at stake, or there is a transformation of the markup into a DTABf markup.

When a new structural phenomenon is encountered there usually is a semantically equivalent tagging solution already provided by the DTABf. The facsimile in example 1 represents a case where the (discontinuous) quotation is presented inline whereas the bibliographic citation is given in the margin. This markup can be transformed into a DTABf solution where discontinuous quotation parts, the linear order of the text and the correct bibliographic references are handled.



Pſal. 139.

nicht ſagen / das Chriſtus allein im Himmel vnd nicht bey vns
auff Erden ſey / ſintemahl er ſitzet zu der Rechten der krafft Got-
tes / welche ſich nicht theilen leſt / ſondern allenthalben iſt / vnd alles
erfüllet. Denn es ſtehet geſchrieben im 139 Pſalm. Wo
ſoll ich hingehn für deinem Geiſt / vnd wo ſol ich hinfliehen für
deinem Angeſicht? Fuhre ich gehn Himmel / ſo biſtu da / bettet
ich mir in die Helle / ſiehe ſo biſtu auch da / Nehme ich flügel der
Morgenröte / vnd bliebe am euſſerſten Meer / ſo würde mich doch
deine Handt daſelbſt führen / vnd deine Rechte mich halten: vnd

```
nicht ſagen/ das Chriſtus allein im Himmel vnd nicht bey vns<lb/>
<cit xml:id="bibl9" next="#quote12">
  <bibl>
    <note place="left">Pſal. 139.</note>
  </bibl>
</cit>
auff Erden ſey/ ſintemahl er ſitzet zu der Rechten der krafft Got-<lb/>
tes/ welche ſich nicht theilen leſt/ ſondern allenthalben iſt/ vnd alles<lb/>
erfüllet. Deñ alſo ſtehet geſchrieben im 139 Pſalm.
<cit xml:id="quote12" prev="#bibl9">
  <quote>Wo<lb/>
    ſoll ich hingehn für deinem Geiſt/ vnd wo ſol ich hinfliehen für<lb/>
    deinem Angeſicht? Fuhre ich gehn Himmel/ ſo biſtu da/ bettet<lb/>
    ich mir in die Helle/ ſiehe ſo biſtu auch da/ Nehme ich flůgel der<lb/>
    Morgenrôte/ vnd bliebe am euſſerſten Meer/ ſo würde mich doch<lb/>
    deine Handt daſelbſt führen/ vnd deine Rechte mich halten:</quote>
</cit>
```

Example 1: Discontinuous Quotations

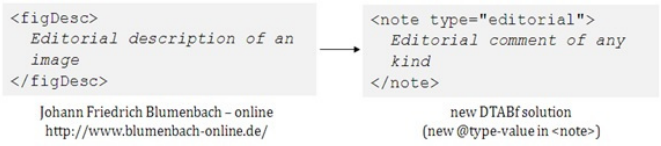
Texts from external projects can contain markup that is not part of the DTABf format. In many of these cases the original tagging there is a straightforward transformation into an already existing DTABf solution. This case is illustrated in example 2 where <unclear>-element is replaced by the <gap>-element which is part of DTABf.

Example 2: Tagging of text loss

New Phenomena that Require Changes to the DTABf

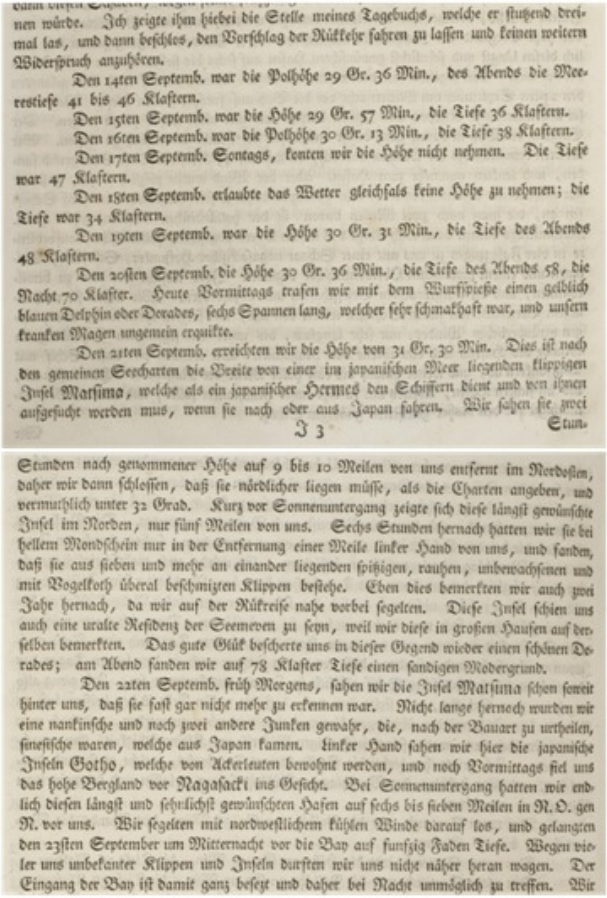
Changes to the DTABf are carried out only if they are consistent with the existing tag set and do not introduce ambiguities to the format. Changes concern mainly attributes or values and less frequently TEI elements or modules. Possible scenarios for cases where the new requirements cannot be handled within the DTABf are the following:

- New texts may contain structures which are new to the DTABf, e.g. due to a new text type or document type (e.g. manuscripts).
- The structural depth upon which the external text has been annotated has no equivalent within the DTABf. Example 3 illustrates that case: a new attribute-value-pair (@type="editorial") has been introduced into the DTABf to cope with editorial descriptions of an image.



Example 3: editorial comments in notes

- Gaps in the documentation can lead to uncertainties about the markup-elements to be applied.



nen würde. Ich zeigte ihm hiebei die Stelle meines Tagebuchs, welche er stutzend dreimal las, und dann beschlos, den Vorschlag der Rükkehr fahren zu lassen und keinen weitern Widerspruch anzuhören.

Den 14ten Septemb. war die Polhöhe 29 Gr. 36 Min., des Abends die Meerestiefe 41 bis 46 Klaftern.

Den 15ten Septemb. war die Höhe 29 Gr. 57 Min., die Tiefe 36 Klaftern.

Den 16ten Septemb. war die Polhöhe 30 Gr. 13 Min., die Tiefe 38 Klaftern.

Den 17ten Septemb. Sontags, konten wir die Höhe nicht nehmen. Die Tiefe war 47 Klaftern.

Den 18ten Septemb. erlaubte das Wetter gleichfals keine Höhe zu nehmen; die Tiefe war 34 Klaftern.

Den 19ten Septemb. war die Höhe 30 Gr. 31 Min., die Tiefe des Abends 48 Klaftern.

Den 20sten Septemb. die Höhe 30 Gr. 36 Min., die Tiefe des Abends 58, die Nacht 70 Klafter. Heute Vormittags trafen wir mit dem Wurfspieße einen gelblich blauen Delphin oder Dorades, sechs Spannen lang, welcher sehr schmakhaft war, und unsern kranken Magen ungemein erquikte.

Den 21ten Septemb. erreichten wir die Höhe von 31 Gr. 30 Min. Dies ist nach den gemeinen Seecharten die Breite von einer im japanischen Meer liegenden klippigen Insel Matsima, welche als ein japanischer Hermes den Schiffern dient und von ihnen aufgesucht werden mus, wenn sie nach oder aus Japan fahren. Wir sahen sie zwei

Stunden nach genommener Höhe auf 9 bis 10 Meilen von uns entfernt im Nordosten, daher wir dann schlossen, daß sie nördlicher liegen müsse, als die Charten angeben, und vermuthlich unter 32 Grad. Kurz vor Sonnenuntergang zeigte sich diese längst gewünschte Insel im Norden, nur fünf Meilen von uns. Sechs Stunden hernach hatten wir sie bei hellem Mondschein nur in der Entfernung einer Meile linker Hand von uns, und fanden, daß sie aus sieben und mehr an einander liegenden spitzigen, rauhen, unbewachsenen und mit Vogelkoth überal beschmizten Klippen bestehe. Eben dies bemerkten wir auch zwei Jahr hernach, da wir auf der Rükreise nahe vorbei segelten. Diese Insel schien uns auch eine uralte Residenz der Seemeven zu seyn, weil wir diese in großen Haufen auf derselben bemerkten. Das gute Glük bescherte uns in dieser Gegend wieder einen schönen Derades; am Abend fanden wir auf 78 Klafter Tiefe einen sandigen Moddergrund.

Den 22ten Septemb. früh Morgens, sahen wir die Insel Matsima schon soweit hinter uns, daß sie fast gar nicht mehr zu erkennen war. Nicht lange hernach wurden wir eine nankinsche und noch zwei andere Junken gewahr, die, nach der Bauart zu urtheilen, sinesische waren, welche aus Japan kamen. Linker Hand sahen wir hier die japanische Inseln Gotho, welche von Ackerleuten bewohnt werden, und noch Vormittags fiel uns das hohe Bergland vor Nagasacki ins Gesicht. Bei Sonnenuntergang hatten wir endlich diesen längst und sehnlichst gewünschten Hafen auf sechs bis sieben Meilen in N. O. gen N. vor uns. Wir segelten mit nordwestlichem kühlen Winde darauf los, und gelangten den 23sten September um Mitternacht vor die Bay auf funfzig Faden Tiefe. Wegen vieler uns unbekanter Klippen und Inseln durften wir uns nicht näher heran wagen. Der Eingang der Bay ist damit ganz besezt und daher bei Nacht unmöglich zu treffen. Wir

E. Kaempfer, Geschichte und Beschreibung von Japan, vol. 1, 1777, p. 69/70

Example 4: Encoding as list items or as paragraphs?

- New TEI P5 releases may introduce changes to tei_all which may affect the DTABf.

Example 5: @unit vs. @type within
<biblScope> (TEI header) in release 2.3.0

Ensuring the Consistency of DTABf Encoded Texts

DTABf Encoding Levels

With the growth of the DTABf it gets increasingly difficult and time consuming to apply the whole range of possible DTABf annotations to each DTA corpus text individually. Therefore we have introduced three levels of annotation which allow for a quick check concerning the extent of interoperability of a text with other texts of the corpus. Each level corresponds to a set of elements; the element lists of the three levels are disjoint. Level 1 lists all the mandatory elements for DTABf conformity (e.g. , <head>, <p>, <lg>, <figure>, <pb>), level 2 those that are recommended (e.g. <cit>, <opener>, <closer>, <lb>, <hi>), level 3 the optional elements (e.g. <persName>, <placeName>, <foreign>). For example, if a document is DTABf-level 3 conformant, all elements of level 1 to 3 must have been applied exhaustively for any element according to the DTA encoding documentation. If elements are only partially applied (e.g. partial application of <persName>), the document is not level-3-conformant and thus not interoperable on that level.

Training and Tools for Users

The existence of a comprehensive documentation is a necessary prerequisite for the applicability of the DTABf by a larger user community. In addition, the DTA offers workshops and tutorials where users learn to apply the DTABf in a consistent way.

Furthermore, text edition according to the DTABf is supported by DTAoX, a framework for the Author mode of the oXygen XML Editor. DTAoX provides an ad hoc visualization of DTABf tagged text passages, of the

annotation levels they belong to, and potential discrepancies from the DTABf.

Conclusion and Further Work

For the supervised growth of the DTABf we make extensive use of the wide range of customization mechanisms the ODD provides. We plan to include schematron rules that will enable us to formulate a higher expressiveness of restrictions. For example, we would like to restrict the usage of some elements which may be used within the <teiHeader>- or <text>-area (e.g. <msDesc>) to either the one or the other which is currently not possible in the ODD mechanism itself.
DTABf currently serves as best practice format for the encoding of historical printed texts in the CLARIN-D project ([CLARIN-D User Guide]). For a better visibility of the DTABf we plan to publish a CMDI-profile of the DTABf-metadata on the CLARIN-EU level where the DTABf metadata elements and attributes are connected to the ISOcat registry as well as conversion routines for the transformation of DTABf conformant header metadata into CMDI. With these efforts, we want to ensure the further, long-term maintenance and lifecycle of the DTABf beyond the duration of the DTA project.

Bibliography

[1] Geyken, Alexander; Haaf, Susanne and Wiegand, Frank (2012): The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora. In Proceedings of Konvens. Wien, 2012, pp. 383-391. [online version]

[2] Bryan Jurish (2010): More than Words: Using Token Context to Improve Canonicalization of Historical German. JLCL 25(1): 23-39. [online version]

For further references [http://www.deutschestextarchiv.de/doku/publikationen].

Links

- Blumenbach – online: http://www.blumenbach-online.de/
- CLARIN-D: http://www.clarin-d.de/

- CLARIN-D User Guide: http://clarin-d.de/en/language-resources/userguide.html
- DDC: http://www.deutschestextarchiv.de/doku/software#ddc
- DTA: http://www.deutschestextarchiv.de/
- DTAE: http://www.deutschestextarchiv.de/dtae/
- – English: http://www.deutschestextarchiv.de/doku/basisformat_table?lang=en
  – German: http://www.deutschestextarchiv.de/doku/basisformat/
- DTABf ODD: http://www.deutschestextarchiv.de/basisformat.odd
- DTABf RNG: http://www.deutschestextarchiv.de/basisformat.rng
- Polytechnisches Journal: http://www.polytechnischesjournal.de/
- TEI P5 guidelines, ch. 15.5: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCREC
- TEI P5 version 2.3.0 release notes: http://www.tei-c.org/release/doc/tei-p5-doc/readme-2.3.0.html

(All URLs cited in this paper were retrieved: 2013-06-21.)

Promoting the linguistic diversity of TEI in the Maghreb and the Arab region

Hudrisier, Henri; Zghibi, Rachid; Sghidi, Sihem; Ben Henda, Mokhtar

Presentation

Since many centuries, the Maghreb region is experiencing significant linguistic hybridization that slowly impacts on its cultural heritage. Besides Libyan, Latin and Ottoman contributions, significant other amounts of resources in various cultures and languages have been accumulated in the Maghreb region, either derived from classical Arabic (i.e. regional dialects) or from various dialects of Berber (i.e.

Kabyle). Several resources are even composed simultaneously in several common or restricted languages (literary Arabic, colloquial Arabic, French, English, Berber) like newspapers, “city printing”, advertising media, popular literature, tales, manuals for learning languages, etc. These resources are often written in a hybrid script mixing both classical and vernacular Arabic, or combining transliteration forms between Latin, Arabic and Tifinagh (traditional Berber script). Unlike many traditional textual resources (conventional printed documents and medieval manuscripts), it does not exist today vast corpora of texts in vernacular idioms and scripts. But our hypothesis is that the growing awareness of the diversity of these textual resources would rapidly result in an exponential increase of the number of researchers interested in collecting and studying classical old texts and oral resources. The standard TEI encoding format provides in this respect a unique opportunity to optimize these resources by ensuring their integration into the international cultural heritage and their use with maximum technical flexibility. The "HumanitéDigitMaghreb" project, which is the subject of this intervention, intents to address several aspects of theses research objectives and to initiate their appropriation.

Research hypothesis

The project targets both oral corpus and the rich text resources written in the Maghreb region. It focuses particularly on the continuity, for more than 12 centuries, of a classical still alive Arabic language and on the extreme hybridization of vernacular languages sustained by the rich Libyan, Roman, Hebrew and Ottoman influences and by the more recent French, Spanish and Italian linguistic interference. In short, the Maghreb is a place of extremely abundant, but much unexploited, textual studies.
Our project permits comparative visions to understand how to transform TEI originally designed for classical and modern European languages (Latin, medieval languages, etc. ...) in order to work on corpora in literary Arabic and in mixed languages and scripts. For example, how researchers from the Maghreb, who invest in the French metric study and fully understand the TEI markup, can understand the subtlety of Arabic meter markup? How do they develop and give examples, when possible, of markup terminological equivalents of metric description in English,

French and Arabic? How can they see if there are really specific «Arabic» structural concepts and then provide the appropriate tags for them. These questions can concern “manuscripts”, “critical apparatus”, “performance text", etc...? For “TEI speech”, we assume, however, that it is not really likely to be the specific method to apply although much work remains to be done. Doing this, we are aware that researches on similar adaptations are undertaken in other languages and cultures: Korean, Chinese, Japanese ... Theses adaptations and appropriations of the TEI experiences are of high interst for us.

Core questions

As a starting point, we consider that the use of TEI in the Maghreb and the Middle East is still sporadic and unrelated. The existing work is mainly concentrated on the study of manuscripts and rare books. This focus can be explained primarily by the existence of large collections of Oriental manuscripts in western digital collections that are TEI encoded since a long time. It can also be explained by the urgency felt within the Arab cultural institutions to accelerate the preservation of cultural heritage from deterioration. Thus, we assume that TEI relatively profited from all experiences and projects for encoding Arabic manuscripts. However, this effort seemingly still needs a larger amount of feedbacks of other nature, generated from other types of resources with other forms of complexity (mainly linguistic and structural). The question that drives us here is to know how the complexity of that cultural heritage (that of the Maghreb as much as we are concerned) would be of any contribution to TEI? How to define its cultural and technological distinctiveness compared to the actual TEI-P5 and what are the solutions?

Methodology

In the project "HumanitéDigitMaghreb", we particularly focus on the methods of implementing the TEI to address specific complex structures of multilingual corpus. We achieved some results, but on the long term, we especially concentrate on practical and prospective issues of very large standardized and linguistically structured corpora that will allow, for all linguistic communities (and we concentrate here on the Maghreb world), to constitute appropriate references in order to interact correctly with

translation technologies and e-semantics in the future. On this last point, it is essential that the community of Arab and Berber researchers mobilize without delay to provide these languages (both written and oral) with their digital modernity. Three steps are to be taken in this respect:

1. The first step, which is beyond the limits of our project "HumanitéDigitMaghreb", inevitably involves a linguistic and sociocultural analysis of the Arabic context in order to clarify three points: first, how the TEI, in its current and future versions, would encode the Arab cultural heritage; second, how the Arabic context surpasses the limits of one level of standard cataloging (MARC, ISBD, AACR2, Dublin Core) ; and third, how it succeeds to standardize the different approaches of its heritage scholarly reading.

In its constant evolution, and the need to strengthen its internationalization, the TEI community would undoubtedly profit from these cultural and linguistic characteristics. This would require also that this community be well organized to provide adequate encoding standardized formats for a wide range of linguistically-heterogeneous textual data. We can imagine here the encoding needs of electronic texts in Arabic dialects profoundly scattered with transliterated incises or written in different characters. These texts are potentially very complex. Besides connecting these materials to each other, like in parallel data (often bilingual), there are further levels of complexity inherent to the use of character sets and multiple non-standard transcription systems (different from the International Phonetic Alphabet) and related to the need of transcribing the speech in an overwhelmingly oral society, which poses interesting encoding problems.

2. The second step, which is under the scope of our proposal, is to produce TEI standard references in local languages and to introduce them to academic and professional communities. These standards help address issues of specific linguistic complexity like hybridization of digital resources (local dialects) and preservation of a millenary oral and artistic heritage. Thus, the issue of character sets is not without consequence to represent local dialects, in large part because many of their cultural aspects were not taken into account in the development of existing standards (transcribing numbers and symbols, some forms of ligatures,

diplomatic and former alphabets). There are, for example, many properties of the Arabic or Berber languages, as the tonal properties, regional synonymy and classical vocalization, (notarial writing) that require special treatment. Current standards, in particular the Unicode and furthermore ISO 8859 standards, do not take into account many of these aspects.
3. The third step, in which we are also engaged, is the creation of a community of practice specialized in the treatment of specific resources. We note here that most of these resources are potentially complex and certain features require probably specific markup arrangements. This means that a dynamic environment is required to specify the encoding of these documents - an environment in which it is easy to encode simple structures, but where more complex structures can be also encoded. Therefore, it is important to have specifications that can be easily extended when new and interesting features are identified.
We are interested in TEI not only for its collegial dynamics open on non-European linguistic diversity (Japan, China, Korea…), but also for its eclectic research disciplines (literature, manuscripts, oral corpus, research in arts, linguistics...) and its rigor to maintain, enrich and document open guidelines on diversity ensuring at the same time the interoperability of all produced resources.

Results

The results of our work are reflected through a website that lists a collection of TEI encoded samples of resources in areas such as music, Arabic poetry, Kabyle storytelling and oral corpus. To achieve this, we went through a fairly rapid first phase of TEI guidelines appropriation. The second phase would be a larger spreading of the TEI guidelines among a wider community of users including graduate students and mostly scholars not yet convinced of the TEI added-value in the Maghreb region. Those could be specialists of Arabic poetry, specialists of the Berber language, musicologists, storytelling specialists... The translation of the TEI P5 in French and Arabic, but also the development of a sample corpus and the construction of TEI multilingual terminology or glossary in English/French/Arabic, seems very necessary.
We also intend to propose research activities within other communities acting at national and regional levels in order to be in total synergy

with the international dynamics of TEI. We have been yet involved in an international project, the "Bibliothèque Numérique Franco-Berbère" aimed at producing Franco-Berber digital resources with a funding from the French speaking International organization. In short, by getting engaged in the school of thought of Digital Humanities and TEI, we explicitly intend to give not only a tangible and digital reality to our work, but we try to make it easily cumulative, upgradable and exchangeable worldwide. More specifically, we expect that our work be easily exchangeable between us and our three Maghreb partner languages (Arabic, French, Berber) beside English.

Apart from the emerging issue of management and setting a standardized and interoperable digital heritage, it is obvious that specialists in this literary heritage should largely explore the methods of study and cataloging. Therefore, this article is limited to discuss only questions of scholars and professionals (libraries and research centers) appropriation of digital humanities tools and services in the Oriental context. We will focus, among other issues, on compared cultural problems by facing European ancient manuscripts study to the Arabic cultural context.

Bibliography

- ABBÈS R. (2000). "Encodage des corpus de textes arabes en conformité à la TEI, outils et démarche technique". Rapport final de projet DIINAR-MBC.
- Bauden F., Cortese Delia Ismaili, and other (2002). Arabic Manuscripts. A Descriptive Catalogue of Manuscripts in the Library of The Institute of Ismaili Studies.
- Burnard, L. (2012). "Encoder l'oral en TEI#: démarches, avantages, défis…. Présenté à Conférence à la Bibliothèque Nationale de France, Paris: Abigaël Pesses.
- Guesdon, Marie-Genvière (2008). "Bibliothèque nationale de France: Manuscripts catalogue 'Archives et manuscrits'". Paper presented at the Fourth Islamic Manuscript Conference, Cambridge
- Hall, G. (2011). Oxford, Cambridge Islamic manuscripts catalogue online. http://www.jisc.ac.uk/whatwedo/programmes/digitisation/islamdigi/islamoxbridge.aspx

- Henshaw, C.(2010). "The Wellcome Arabic Manuscript Cataloguing Partnership", in: News in brief, D-Lib Magazine, March/Apri. http://www.dlib.org/dlib/march10/03inbrief.html
- Ide, N. (1996). “Representation schemes for language data: the Text Encoding Initiative and its potential impact for encoding African languages”. In CARI’96
- Ide, N. M., Véronis, J. (1995). Text Encoding Initiative: Background and Contexts. Springer.
- Jungen, C. (2012). “Quand le texte se fait matière”. Terrain, n° 59(2), 104#119.
- Mohammed Ourabah, S., Hassoun, M. (2012). “A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts”. Journal of the Text Encoding Initiative,
- Pierazzo, E. (2010). “On the Arabic ENRICH schema”. Wellcome Library Blog, 27 August, http://wellcomelibrary.blogspot.com/2010/08/guest-post-elena-pierazzo-on-arabic.html
- Véronis, J. (2000). Parallel Text Processing: Alignment and Use of Translation Corpora. Springer.

XQuerying the medieval Dubrovnik

Jovanović, Neven

To anyone with the time and patience to study the voluminous Acta consiliorum [of Dubrovnik / Ragusa], wrote Fernand Braudel in 1949, they afford an opportunity to observe the extraordinarily well-preserved spectacle of a medieval town in action. The archival series of decisions and deliberations made by the three administrative councils of Dubrovnik consist of hundreds of handwritten volumes, predominantly in Latin and still not published in its entirety, spanning the period from 1301 until 1808

(the year the Republic of Ragusa was abolished by Napoleon's Marshal Auguste de Marmont) [1].

In collaboration with Croatian Academy of Sciences and Arts, Institute of Historical Sciences - Dubrovnik, which is the current publisher of the series Monumenta historica Ragusina (MHR), we have undertaken a pilot project of converting to TEI XML the Volume 6 of MHR. The volume publishes the so-called Reformationes of Dubrovnik councils from the years 1390-1392; it was edited by Nella Lonza and Zdravko Šundrica in 2005 [2]. In this text, different salient points of the Reformationes (meetings, names of persons and places, dates, values and measures, themes, textual annotations) are being marked and the markup decisions are carefully documented, all with the twofold intention of, first, enabling XQuery searches of the Reformationes through the BaseX database [3] not just by us, but by other users, and, second, preparing the documentation for further encoding of other MHR volumes (producing of a "MHR in XML" data set we see as a necessary, but necessarily extensive task).

The small city of Dubrovnik and its relatively closed, but well-documented society were already subjected to a database-driven research project, carried out in 2000 by David Rheubottom (then at the University of Manchester), who used archival records to examine the relationship between kinship, marriage, and political change in Dubrovnik's elite over a fifty-year period, from 1440 to 1490 [4]. But where Rheubottom, relying on classical relational database, extracted records from original text, abstracting data from words [5], we intend to use the advantages of XML to interpret not only data, but its relationship with the words (enabling also research of e. g. the administrative formulaic language). Where Rheubottom built his database to explore one set of problems over a limited time series, we intend to make it possible for different researchers to pursue their different interests in the framework which could, eventually, embrace all recorded decisions from 500 years of Dubrovnik's history. Last but not least, Rheubottom's database remained unpublished -- his interpretations were published as a printed book; today we have the possibility to publish (or, to open access to) not only the TEI XML annotated version of the MHR 6, but also the documentation of our encoding principles, as well as the XQueries which we find useful or

interesting. Publishing the XQueries makes our research repeatable and reproducible [6]; presenting them in a graded, logically organized way, from the simplest and easiest to more complex and difficult, ensures their educational value.

The TEI XML encoding standard is sometimes criticized for its "there's more than one way to do it" approach. We hope to show that what one person regards as a drawback, the other can regard an asset; we hope to demonstrate not only how we chose among available TEI elements and attributes to solve specific encoding challenges (e. g. to encode commodity prices, persons referred to also by their father's name, absence of explicit dates in datable documents, election results), but also to show the ongoing process of documenting the selected combinations and their "constellations", both in the free prose, more accessible to laypersons, and in the format of XML Schema Documentation of the TEI subset produced by encoding [7].

XQuery is a powerful and expressive programming language, but it is certainly not something that common computer users normally see; by and large, the XQuery layer remains hidden and only selected, prefabricated queries get displayed. Mastering XQuery to explore a database can seem a daunting task, and one best left to non-academic specialists. But let us not forget that the historians who plan to explore records of medieval Dubrovnik in their existing form have already shown enough motivation to master a similarly daunting accessory task of learning medieval Latin (and, in some cases, medieval palaeography). Also, looking at a resource such as The Programming Historian collaborative textbook [8], one can see to what computing depths some historians are prepared to go to be able to pose interesting questions to their material. The ideal user of the MHR in XML is an algorithmically literate medieval scholar, one which does not consider computers as black boxes; perhaps the MHR in XML can itself produce, that is educate, such digital humanists. Because, as Aristotle wrote, Anything that we have to learn to do we learn by the actual doing of it.

Bibliography

- [1] Croatian State Archive in Dubrovnik, "Pregled fondova i zbirki, A.1.5. Dubrovačka Republika do 1808." ["A list of archival series

and collections, A.1.5 The Republic of Dubrovnik until 1808"], http://www.dad.hr/fondovi_zbirke.php.

- [2] Lonza, Nella and Šundrica, Zdravko (eds). Odluke dubrovačkih vijeća 1390-1392 [Deliberations of the Councils of Dubrovnik 1390-1392]. Dubrovnik: HAZU, Zavod za povijesne znanosti u Dubrovniku, 2005.
- [3] 'BaseX. The XML Database', http://basex.org/
- [4] Rheubottom, David. Age, Marriage, and Politics in Fifteenth-Century Ragusa. New York, Oxford University Press, 2000.
- [5] Rheubottom, David, 'Computers and the political structure of a fifteenth-century city-state (Ragusa)', in History and Computing, edited by Peter Denley, Deian Hopkin, Manchester University Press, 1987, pp. 126–132.
- [6] 'BaseX Adventures', http://www.ffzg.unizg.hr/klafil/dokuwiki/doku.php/z:basex-adv.
- [7] 'Reformationes consiliorum civitatis Ragusii: encoding guidelines', http://www.ffzg.unizg.hr/klafil/dokuwiki/doku.php/z:dubrovnik-reformationes [under construction]
- [8] Crymble, Adam et al. 'The Programming Historian 2', http://programminghistorian.org/

Analyzing TEI encoded texts with the TXM platform

Lavrentiev, Alexei; Heiden, Serge; Decorde, Matthieu

TXM (http://sf.net/projects/txm) is an open-source software platform providing tools for qualitative and quantitative content analysis of text corpora. It implements the textometric (formerly lexicometric) methods developed in France since the 1980s, as well as generally used tools of corpus search and statistical text analysis (Heiden 2010).

TXM uses a TEI extension called "XML-TXM" as its native format for storing tokenized and annotated with NLP tools corpora source texts (http://sourceforge.net/apps/mediawiki/txm/index.php?title=XML-TXM). The capacity to import and correctly analyze TEI encoded texts was one of the features requested in the original design of the platform.

However, the flexibility of the TEI framework (which is its force) and the variety of encoding practices make it virtually impossible to work out a universal strategy for building a properly structured corpus (i.e. compatible with the data model of the search and analysis engines) out of an arbitrary TEI encoded text or group of texts. It should nevertheless be possible to define a subset of TEI elements that would be correctly interpreted during the various stages of the corpus import process (for example, the TEI-lite tag set), to specify the minimum requirements to the document structure and to suggest a mechanism for customization. This work is being progressively carried out by the TXM development team, but it can hardly be successful without an input from the TEI community.

The goal of this paper is to present the way TXM currently deals with importing TEI encoded corpora and to discuss the ways to improve this process by interpreting TEI elements in terms of the TXM data model.

At present, TXM includes an "XML-TEI-BFM" import module developed for the texts of the Base de Français Médiéval (BFM) Old French corpus (http://txm.bfm-corpus.org) marked up according to the project specific TEI customization and guidelines (Guillot et al. 2010). With some adaptation, this module works correctly for a number of other TEI encoding schemas used by several projects: Perseus (http://www.perseus.tufts.edu/hopper), TextGrid (http://www.textgrid.de/en), PUC/Cléo (http://www.unicaen.fr/recherche/mrsh/document_numerique/outils), Frantext (http://www.frantext.fr), BVH (http://www.bvh.univ-tours.fr), etc. However, the use of tags that are not included in the BFM customization and the non respect of some particular constraints (such as a technique of tagging parts of words and of using strong punctuation within the editorial markup elements) may result in lower quality of the TXM corpus (e.g. errors in word counts, collocation analysis or inconvenient display of texts for reading) or even in a failure of the import process due to the limits of the tokenizer used in this module.

A more generic "XML/w+CSV" module allows importing any XML documents (not necessarily TEI) with the possibility to pre-annotate all or selected words using a <w> tag with an arbitrary set of attributes. This module is more robust in terms of producing a searchable corpus but it does not make any use of the semantics of TEI markup. For instance, no difference is made between the text and the header, the notes and variant encodings of the same text segment are all included in the text flow.

To improve the quality of the resulting corpus, it is necessary to "translate" the TEI markup into the various data categories relevant for the TXM data model. This model is relatively straightforward and relies to a large extent on that of the CWB CQP search engine (http://cwb.sourceforge.net). We have already presented the relevant data categories in some detail at the 2012 TEI Members Meeting (Heiden & Lavrentiev 2012) but this time we would like to adopt a more pragmatic approach related to the development of the TXM-TEI import modules.

A corpus is composed of a number of "text units" associated with a set of metadata used mainly to split the corpus in different ways and to perform contrastive analyses. A simple TEI file with one <text> element corresponds usually to a TXM text unit, and the useful metadata can be extracted from the <teiHeader> (or, alternatively, from a separate CSV table).

The second basic element of the TXM data model is the "lexical unit" (or the token), which may be a word or a punctuation mark carrying a number of properties (annotations) inherited from the source document (e.g. the language or a variant form) or generated during the import process (e.g. morphosyntactic description or a lemma suggested by an NLP tool). The properties of the lexical units can be easily searched and analyzed using the CQP search engine. TXM can import a corpus with pre-tagged lexical units but in most cases the tokenization is performed during the import process. In the latter case, it is necessary to pay special attention to the tags that may occur "inside" the tokens. These are typically line or page breaks, or some editorial markup (abbreviation marks, supplied letters, etc.). As far as the milestone-like empty elements are concerned, the TEI has recently adopted a general mechanism using the "break" attribute. As for the word-internal elements with textual content, it is recommended to

pre-tag the words containing such elements using the <w> element before the import process.

The third element of the TXM data model is the intermediate structure of the text which can include sentences, paragraphs, divisions or any other continuously or sporadically marked up text segments. They are represented as XML elements, so proper nesting is required. They can be annotated by properties that can be used in a way similar to the text unit metadata. Intermediate structures can be used to separate "text planes" (such as titles vs. text body, direct speech of various characters in a drama, etc.). Although TXM is not designed for managing various readings in critical editions or stages of text evolution, the mechanism of text planes can be used to analyze and compare different text states or variants.

In the simplest case, a text can be represented as a chain of lexical units. This point of view is by all means relevant for word counts, collocation search and analysis, etc. If the source document contains editorial notes or variant encodings of the same text segment (using <choice> or <app> mechanisms), it is necessary to treat them in one of the following ways:

- eliminate them completely from the search indexes;

- create a separate "text plane" for them and possibly relocate them to special text units or divisions;

- project variant readings as additional "properties" onto the lexical units of the main text chain.

The last but not the least aspect of the import process is building "editions" of corpus texts for convenient reading and displaying extended contexts of the search hits. This is where the rich TEI markup and the know-how of producing fancy-styled outputs may be particularly valuable. The objective is to make it possible to use a set of custom stylesheets (like those developed by Sebastian Ratz ones for the TEI consortium) to render these editions but this requires some further development to ensure compatibility with TXM's features of highlighting search hits and displaying properties of the lexical units. An intermediate solution is currently being experimented to allow the customization of the rendering of selected elements via the CSS class pointing mechanism.

The TXM team is interested in the feedback from any TEI projects willing to analyze their data with the TXM platform and is open to discussion on the improvement of the import modules and their documentation.

Bibliography

- Guillot, C., Heiden, S., Lavrentiev, A., Bertrand, L. (2010). Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval, Lyon, Équipe BFM <http://bfm.ens-lyon.fr/article.php3?id_article=158>.
- Heiden, S. (2010). "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." 24th Pacific Asia Conference on Language, Information and Computation. Éd. Kiyoshi Ishikawa Ryo Otoguro. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010. 389-398. <http://halshs.archives-ouvertes.fr/halshs-00549764>.
- Heiden, S. & Lavrentiev, A. (2012). "Constructing Analytic Data Categories for Corpus Analysis from TEI encoded sources." TEI Conference 2012. College Station, TX, 7-10 November 2012. <http://idhmc.tamu.edu/teiconference/program/papers>.

"Texte" versus "Document". Sur le platonisme dans les humanités numériques et sur la maïeutique TEI des textes ("Text" versus "Document". Platonism in DH and the maieutics of the text)

Miskiewicz, Wioletta

Dans mon intervention j'aimerais partager les réflexions qui se sont imposées à moi, en tant que philosophe qui dirige le site des archives philosophiques (XX siècle) et qui pratique la TEI en lien avec ces archives.

Grâce à sa dimension sémantique, TEI occupe à plus d'un titre une place privilégiée dans le paysage DH. Ainsi l'encodage TEI est un exemple de la coopération homme/machine qui ne se limite pas à l'utilité technologique (telle que la sauvegarde du patrimoine, la rationalisation du traitement et d'accès aux très grands corpus ou encore la simplification de la publication "à la carte"). L'encodage TEI est aussi créatif et il ouvre la voie d'accès à des contenus nouveaux et insoupçonnables avant. C'est la cas par exemple pour les diverses visualisations des contenus, pour les analyses stylométriques, scientométriques, etc. Et enfin, TEI révèle aussi certaines vérités sur la nature des objets de recherches dans le domaine SHS.

Une tension est palpable au sein de la TEI. En gros, on peux dire que c'est une tension entre l'encodage linéaire d'une succession des unités linguistiques fixées sur un support "transparent" d'un côté et l'encodage génétique, qui vise à rendre le temps originaire de la production du contenu intelligible, de l'autre.

La TEI a été créée pour ce premier et c'est pourquoi "texte" figure d'une manière programmatique déjà dans son intitulé. Cependant depuis des années la recherche sur l'encodage des manuscrits de travail des écrivains (Flaubert, Proust) est engagée (avec des fortunes diverses). Pour cette approche TEI génétique le "document" prend de plus en plus d'importance. Dans mon travail de chercheur je suis intéressée par les deux tendances (pour des raisons différentes).

Quel est le statut ontique du texte? Quelle sorte d'objet est le texte et de quelle façon existe-t-il? Le texte correspond avant tout à une "surface" perceptible. Même le grand prêtre de la postmodernité, Barthes (dans l'*Encyclopedia Universalis*), en convient. Barthes attribue au texte avant tout une fonction de sauvegarde: "d'une part, la stabilité, la permanence de l'inscription, destinée à corriger la fragilité et l'imprécision de la mémoire; et d'autre part la légalité de la lettre, trace irrécusable, indélébile, pense-t-on, du sens que l'auteur de l'oeuvre y a intentionnellement déposé". Cette fonction de sauvegarde est fondamentalement liée au support matériel et à ses propriétés. Peut-on, dans cette fonction de sauvegarde, limiter la "surface perceptible" du texte aux seules combinaisons des lettres? P. Caton - nous y reviendrons plus tard - montre que certainement pas.

La question de la légitimité du "texte" en circulation par rapport à l'oeuvre de son créateur est vieille comme la communication indirecte. Dans le cas de l'écrit, elle se focalise sur l'intention de l'auteur et elle craint l'éditeur malveillant.

Depuis l'invention de l'imprimerie, l'évolution va dans le sens d'une progression vers l'abstraction, vers la suppression des contenus contextuels liés à la matérialité du texte. L'imprimerie a imposé le règne du texte établi (Scholarly Print Edition) et *de facto* indépendant de son support matériel d'origine à savoir le manuscrit d'auteur. En raison de la popularisation du livre ainsi que de l'impératif de diminuer son prix, nous assistons progressivement aux "dégraissages" du document, à la réduction au strict minimum des informations dont le document d'origine est porteur. Nous assistons au triomphe du texte "pur" dans le minimalisme des éditions de poche et encore plus sur les tablettes. Par ailleurs on peut constater alors qu'au sens strict du terme, les créateurs qui pendant des siècles nous laissaient généralement les manuscrits, puis les tapuscrits, produisent aujourd'hui les fichiers électronique. En ce sens on peut dire, que pour la première fois dans l'histoire de l'humanité ils produisent les "textes".

Mais au fur et à mesure de l'avancement de notre aire numérique et étant donné que l'encodage sémantique TEI vise à représenter avantageusement les sources sur le WEB, la question de la légitimité et de la fidélité aux sources se pose à nouveau et d'une manière plus aiguë. Elle est exacerbée par une fabuleuse augmentation de la quantité des archives *on line* et par le *catch as catch can* omniprésent sur le WEB. Pour nous, c'est l'une des raisons d'aller vers le document encodé TEI qui pourrait devenir garant de la légitimité des sources SHS. TEI pourrait devenir pour les *fichiers sources* ce que fût *Das wohltemperierte Klavier* pour le piano.

Dans cet univers virtuel les questions ontologiques prennent une place centrale. Nous l'avons déjà dit, l'encodage TEI révèle certaines vérités fondamentales sur les relations des chercheurs avec les sources SHS. Notre pratique de la TEI en lien avec les archives, montre que l'analyse

de la *situation d'encodage* [5] peut-être considérée comme l'analyse des intentionnalités à l'oeuvre dans toute lecture possible.

L'idée porteuse des e-archives est de remplacer la consultation matérielle des archives par leurs consultations en ligne. Cela a de nombreux avantages qui justifient le cout élevé de l'entreprise. Dans l'idéal, un lecteur en ligne doit pouvoir accéder à toutes les informations et à tous les contenus des archives d'origine non seulement d'une manière plus commode, mais aussi enrichies par l'expertise du site qui les édite. Dans le cas où le document-source est représenté par un fichier XML/TEI, l'objet virtuel consulté va être inévitablement construit par l'édition électronique. TEI peut faire de cette transformation inévitable un enrichissement. Mais faut-il imposer ici comme norme, que l'objet virtuel consulté ainsi c'est un "texte" au sens traditionnel du terme?

Du point de vue étymologique le mot "texte" vient du latin "*textilis*" et "*textilis*" à son tour de "*textus*" - le participe passé de *textere* - tisser. Donc, le mot "texte" vient de l'activité de tisser, vient de l'action. Mais à la différence de la tapisserie, l'auteur n'est pas le seul à tisser, le lecteur tisse aussi[6]. Comment peuvent-ils avoir tous deux à faire avec le même "objet"? Cette grande question est à l'origine des théories ontologiques et épistémologiques sur le texte et l'identité/permanence de sa signification.

Une longue tradition herméneutique s'occupait de l'explication des intentions des créateurs des oeuvres, de l'explicitation de la signification de l'oeuvre dans son identité absolue (même la déconstruction postmoderne est une étape de cette démarche). La nouveauté radicale de l'encodage TEI dans ce contexte, consiste dans la redécouverte du document lui-même et dans la considération de l'intentionnalité du lecteur avant tout. L'encodeur qui est en contact immédiat avec le document (où au moins de son facsimilé) est l'accoucheur du texte ou des texte(s) possible(s) latent dans le document. L'encodeur doit interroger le document pour l'amener à extérioriser son texte en partant de la

[5] Nous avons analysé le concept de la *situation esthétique* à partir de Roman Ingarden. La situation d'encodage possède une structure ontique en tout point comparable, mais c'est une *situation heuristique*.

[6] R. Ingarden, *The Cognition of the Literary Work of Art*, Illinois:Northwetern University Press, 1973.

matérialité de la source. Il pratique la maïeutique de l'oeuvre et le travail d'encodage révèle la nature essentiellement psychophysique du document, l'importance basique de la perception de son apparence matérielle.

La véritable spécificité de la perception de ces objets particuliers que sont les documents dans l'*attitude de la lecture* est très peu prise en compte dans les théories herméneutiques et dans les théories traditionnelle du texte.

Il a fallu attendre le début du XXème siècle pour qu'une théorie des *processus effectifs* de l'écriture et de la lecture voit le jour. Il s'agit de la théorie des Actions et des Produits (*APT*) de Kazimierz Twardowski. Dans la philosophie la langue est traditionnellement assujettie a exprimer les concepts. Seuls les stoïciens ont pressenti la potentialité du langage d'être un objet *suis generis*. Le langage est considéré comme un reflet direct de la pensée. L'écriture est considédée comme la représentation du langage. La deuxième moitié du XIXème siècle voit enfin naître la théorie de l'intentionnalité (Franz Brentano, 1838 - 1917) qui, pour la première fois, et sur la base d'une psychologie descriptive de la conscience, jette les ponts entre la pensée et le langage. Son élève - Kazimierz Twardowski (1866-1938) affirmera que la langue ne dit pas seulement quelque chose mais aussi sur quelque chose et que même les expressions impossibles ont un objet (par exemple - "le carré rond"). Les bases ontologiques de la sémantique moderne sont ainsi posées. Face au danger logiciste que cette théorie comporte (cf. Lukasiewicz, Lesniewski - ses élèves), Twardowski présente en 1911 la *théorie des actions et des produits* qui est une théorie interdisciplinaire aux confins de la grammaire, de la psychologie et de la logique.

L'homme est selon cette théorie auteur/créateur et produit les objets par le biais de ses actions. En pensant, l'homme peut décider de fixer sa pensée dans l'écriture: l'homme construit alors - dans une langue concrète - des phrases (propositions). Du point de vue ontique, ses pensées en tant que processus psychophysiques concrets, ne sont pas identiques avec leur résultat fixé dans l'écriture concrète. Une fois la proposition couchée sur le papier, l'homme devient le premier lecteur de ses pensées. C'est d'abord en tant que lecteur qu'il les corrige. Le produit de sa cognition est toujours un produit psychophysique, sauf quand - oublié de tous, latent - il attend d'être lu. En attendant il est alors uniquement potentiellement une écriture.

Les traces d'encre sur le papier existent tant qu'elles durent. Elles sont périssables mais autonomes du point de vue ontique. Par contre l'écriture n'est pas autonome du pont de vue ontologique. Elle a besoin d'être ravivée pour devenir ce qu'elle est, à savoir - un produit de la pensée.

La particularité de la perception d'une écriture sur un objet matériel (papyrus, parchemin, papier, etc.) est bien saisissable par ressemblance et par dis-analogie avec une perception possible de ce même objet matériel mais en tant qu'objet d'art plastique, donc dans *l'attitude esthétique*. Imaginons la pierre de Rosette comme un objet décoratif couvert d'ornements répétitifs. Les lettres et les mots (reconnus comme tels ou non) feront partie de la perception en tant qu'éléments d'une perception holistique de l'objet. Les traces matérielles qui correspondent à l'écriture y seront considérées tout d'abord comme les autres traces matérielles, en tant qu'éléments fonctionnels dans le construction de l'objet de l'expérience esthétique. Ils feront partie des aspects[7] par lesquels l'objet esthétique et ses valeurs se présentent, tout d'abord sensoriellement, à celui qui le perçoit.

Par contre dans la perception dans *l'attitude de la lecture*, les aspects perceptifs relèveront avant tout de la signification possible de l'écriture. Cependant ici aussi l'action commence par une impulsion matérielle. Le texte est un objet psychophysique et une théorie des *aspects*, c'est à dire des items sensoriels et perceptifs dans le processus de la construction du texte serait utile.

Est-ce que cela veut dire qu'après la grande époque des herméneutiques qui partaient de l'idée de la signification d'un texte conçu par les intentions de l'auteur, une révolution copernicienne aura lieu grâce à la TEI et on reconnaîtra à l'encodeur le pouvoir constitutionnel par rapport aux textes? Pour répondre à cette question, retraçons les principales conceptions ontologiques du texte. Nous allons distinguer trois types de concepts ontologiques du "texte": la conception platoniste (A), la conception positiviste (B) et finalement la conception sémantique (C).

(A) Dans la SEP nous pouvons lire que le terme: *platonist* signifie dans un sens contemporain: "that there exist such things as abstract objets - where

[7] "Aspects" ne signifie pas : "perspectives", "côtés" ou "fragments". Ce sont plutôt les parcours sensoriels dans la construction de l'objet.

an abstract objetc is an objetct that not exists in space or time and which is therefore entirely non-physical and non-mental"[8].

Cette idée remonte à Platon et à sa métaphore de la caverne: nous ne voyons que les ombres de la vrai réalité, qui comme le soleil dans la métaphore platonicienne se trouve derrière notre dos. Cette théorie a connu les versions plus où mois radicales et la version contemporaine exposée dans la SEP est très modérée. Intuitivement elle est facile à comprendre grâce aux idéalités mathématiques. Les nombres n'existent pas seulement dans tous les actes concrets de dénombrement. Nous reconnaissons sans difficulté la vérité de la phrase: "Il est vrai que les nombres existent". La chose se complique si nous posons la question: est-ce que ce sont les mathématiciens qui ont créé les nombres? La position platoniste consiste à dire: non, les nombres existent indépendamment de l'homme, ils existent à priori et au-delà tout calcul concret, ils n'ont pas été créés par l'homme, ils ont été, éventuellement, d'une certaines façon découverts.

Dans le cas du texte, comme dans celui de l'oeuvre d'art, cette position ontologique est plus nuancée dans la mesure où on reconnaît ici à l'homme plus facilement son pouvoir de créer. Mais une fois l'oeuvre créée il rejoint le royaume apriorique des êtres identiques et durables.

La conception platoniste du texte est omniprésente dans les DH. On peut le voir très bien sur l'exemple de l'ontologie DH proposée par Renear & Dubin.

Renear & Dubin partent dans leurs considérations ontologiques de la typologie FRBR (Functional Requirements for Bibliographic Records/ Spécifications Fonctionnelles de Notices Bibliographiques) de l'IFLA concernant les entités possibles à cataloguer par les bibliothécaires. Dans le premier groupe FRBR on distingue quatre unités: oeuvre (Q), expression (par ex. la traduction de Q par XY), manifestation (une édition de cette traduction chez un éditeur Z) et finalement un item (l'exemplaire que j'ai dans ma bibliothèque). Renear & Dubin démontrent à l'aide du concept de la "propriété rigide", que trois de ces unités ne sont

[8] http://plato.stanford.edu/entries/platonism/. Deux des quatre types FRBF ("oeuvre" et "expression") n'ont aucune réalité psychophysique.

pas des "types" mais uniquement les "rôles" de la première[9]. Même si leur raisonnement est rigoureux et que leurs investigations contiennent énormément d'observations justes, on est obligé de constater que le cadre général de leur raisonnement, à savoir l'affirmation qu'uniquement l'"oeuvre" est un type ontologique, est l'expression d'un pur platonisme. Car, deux des quatre unités du premier groupe FRBR, à savoir l'oeuvre et l'expression sont parfaitement abstraites: aucune expérience immédiate psychophysique n'est ici possible. On peut montrer[10], qu'elles sont des constructions conceptuelles postérieures à toute expérience effectivement possible. Elles sont des constructions conceptuelles utiles pour des besoins de classification (théories ou catalogues) mais elle ne sont pas des moments des expériences possibles. On peux les rencontrer en tant que concepts par le biais de leurs définitions ou par l'abstraction à partir d'une classe de leurs représentants (manifestations et items).

Le seul "type" du premier groupe reconnu par Renear & Dubin "comme type" - est en fait une abstraction!

(B) La *conception positiviste/linguistique* du texte part de la conception du texte en tant qu'unité linguistique. Elle a donc l'allure plus concrète car elle réfère à une connexion unitaire des sens linguistiques structurés. C'est en ce sens que le texte est présent. Dans l'intitulé même du projet TEI: "*Representation of Texts in Digital Form*" et plus précisément "*Encoding Methods for Machine-readable Texts, Chiefly in the Humanities, Social Sciences and Linguistics*".

Dans les TEI Guidelines "text" est un élément du module: "textstructure" avec pour définition: "*text contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, or a corpus sample*"[11].

Dans la pratique concrète de l'encodage "text" arrive après "teiHeader" et contient le contenu intelligible du document à encoder. Le "text" ne contient pas de "metamark" concernant le document même ("contains or describes any kind of graphic or written signal within a document the

[9] Ainsi pour P. Caton (*op. cit.*) le texte est " a matter of contingent social/linguistic circumstances" et les "countable texte" - not a *type* but a *role*.

[10] Par exemple à l'aide de la théorie APT (Actions & Products) de Twardowski.

[11] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-text.html

function of which is to determine how it should be read rather than forming part of the actual content of the document").

Ce concept du "texte" avec lequel opère TEI est en fait emprunté à la fois à la linguistique et à la philosophie. Un texte est une unité structurée des sens langagiers. Idéalement c'est une suite linéaire des propositions (même les expressions bizarres ou "incorrectes" ou autres inventions artistiques peuvent être comprises de cette façon propositionnelle) considérées comme une succession des signes graphiques. Traditionnellement et même dans sa compréhension postmoderne (celle qui inclut les intertextualités, la disparition du texte dans sa déconstruction ou encore la mise sur un pied d'égalité de l'auteur et de l'interprète), ce concept ne réfère que très peu aux propriétés matérielles de ce qui est le véritable support originaire du texte, à savoir le document. En fait, on ne s'y intéresse que s'il est détérioré et que les fragments de l'écriture sont difficilement lisibles, ou encore si ces propriétés matérielles permettent de dater l'écriture ou d'établir l'authenticité du texte et peuvent enrichir le "teiHeader". Le texte transcende ici toujours le *token* qui est son support. Malgré son approche *positiviste* le texte possède finalement une existence unitaire apriorique dans un ailleurs où l'a envoyé l'intentionnalité de son créateur.

Cette compréhension présuppose donc finallement aussi une idée platoniste du texte, car on ne s'intéresse pas ici véritablement aux suites des signes graphiques en tant que fragments du document mais à ce dont ils sont la représentation linguistique et dont l'existence est présupposée a priori: une expression langagière d'une théorie, d'un récit *etc*.

Paul Caton montre bien les limites de cette compréhension du texte. En s'appuyant sur l'analyse des documents, il montre l'importance du contexte "intérieur" du document pour la compréhension du texte. Il démontre sur les "cas extrêmes" (logo, message cryptographié, écriture sur les affiches) l'importance de la fonction que la suite des signes linguistiques remplit dans/sur chaque document. Il met en avant la communication: les suites des signe linguistiques sont non seulement une représentation écrite du langage - ils *communiquent* avant tout. Et pour cette communication, le document donne parfois les informations incontournables. Paul Caton conclut que la distinction tranchée du texte et du contexte dans le document est un artefact.

Cette compréhension le "texte" devient clairement problématique dans le cas d'encodage des manuscrits d'auteur. Les travaux récents d'Elena Pierazzo consacrés à la génétique textuelle montrent l'importance du document pour la reconstruction de son contenu à transmettre. Nonobstant les difficultés que la génétique textuelle pose au niveau de l'affichage, elle révèle l'importance du document et la complexité de l'identité du texte dans/sur le document.

conduit suivre la classification qui obéit au signes graphique sur un support et leur structuration linguistique. Cette observation a une valeur plus générale: la langue et non pas uniquement il faut respecter les exigences objectives

(C) *La conception sémantique du texte*. Notre propre expérience d'encodage a montré, que pour capter le contenu d'un manuscrit philosophique et pour le transmettre pour l'étude à un chercheur future, il ne suffit pas de suivre la classification qui obéit à la structuration linguistique et au positionnement des signes graphiques sur un support. Souvent, pendant l'encodage (et nous allons en montrer quelques exemples pour finir) il faut respecter les exigences *objectives* relevant de la pensée en évolution. Nous nous sommes donc basés sur les propriétés du document autant que sur notre connaissance de la théorie APT de Twardowski.

Notre document à encoder contient une version française de la théorie APT faite par Twardowski lui-même. Ce document n'est pas une traduction proprement dite. Ici Twardowski pense sa théorie en français. Précédemment il l'a déjà formulée en polonais et en allemand.

En partant des caractéristique du document, nous avons pu constater qu'il y avait deux textes correspondant aux deux campagnes d'écritures. Dans notre procédé la langue nous a guidée mais elle ne décidait pas *in fino* sur l'encodage[12]. Ce principe, nous amène vers la conception sémantique du texte.

Sur la troisième voie - sémantique - de la compréhension du concept du texte, le point de départ est toujours donné par la situation d'encodage et par les intentionnalités de lecture que l'encodeur détecte dans le document. Ceci est très proche de la réalité effective de la situation de la lecture et

[12] Cette observation a une valeur plus générale pour les ontologies DH.

de la rencontre (dans l'attitude de lecture) avec l'objet matériel consulté. Ce processus ne présuppose pas l'existence antérieure d'UN texte à reconstituer et il est donc plus ouvert que les deux précédents.
La question se pose à nouveau si, grâce à un tel encodage TEI génétique des manuscrits, ce ne sont plus les textes a priori qui déterminent l'encodage, mais l'encodage qui au fur et mesure de l'avancement de l'encodage donne le texte? Autrement dit: cette troisième voie donne-t-elle à l'encodeur le pouvoir de constituer librement les textes dans son activité effective d'encodage?
Il est vrai: l'encodeur ne travaille pas ici à la re-constitution d'un texte préexistant. Sa liberté est cependant limitée par les déterminations venant du document et de *l'objet théorique que le créateur a fixé dans le document*. Le texte où les textes, arrivent au fur et à mesure de l'encodage: ils n'existent pas avant mais ils ne sont pas constitués d'une manière aléatoire: "The semantic tradition consists of those who believed in the a priori but not in the constitutive powers of mind"[13].
*
L'encodeur TEI est en quelque sorte un Hyperlecteur. Son intentionnalité est celle de tout lecteur possible d'un document et non celle de son créateur. L'encodeur a pour tâche, pour construire le texte, de rendre au mieux les contenus communiqués par le document et de se laisser guider par son objet. Le texte est un produit secondaire d'encodage; il est un a priori non antérieur au travail d'encodage. Ne faudrait-il pas alors, dans l'ordre de l'arbre des documents XML/TEI, remplacer le "texte" par le "document" et réintroduire le "texte" plus tard dans l'embranchement? *This is the question*.

Bibliography

- Burnard, Lou. *Text Encoding for Interchange: A New Consortium*. 2000. [http://www.ariadne.ac.uk/issue24/tei].
- Caton, Paul, *"On the term 'text' in digital humanities", Literary and Linguistic Computing*, vol 28, No.2, 2013, p. 209- 220.
- Crasson, Aurèle and Jean-Daniel Fekete. *Structuration des manuscrits: Du corpus à la région.* Proceedings of CIFED 2004.

[13] J.A. Coffa, *The Semantic Tradition from Kant to Carnap*, Cambridge University Press, 1991, p. 1.

La Rochelle (France), 2004: 162–168. [http://www.lri.fr/~fekete/ps/CrassonFeketeCifed04-final.pdf].

- J.A. Coffa, *The Semantic Tradition from Kant to Carnap*, Cambridge University Press, 1991.
- R. Ingarden, *The Cognition of the Literary Work of Art*, Illinois:Northwetern University Press, 1973.
- W. Miskiewicz, *"La critique du psychologisme et la métaphysique retrouvée - Sur les idées philosophiques du jeune Łukasiewicz"*, Philosophia Scientiae 15/2, – La syllogistique de Łukasiewicz, 2011, p. 21-55.
- W. Miskiewicz, *"Les aspects – Interface entre l'homme et l'œuvre d'art"*, Roman Ingarden: *La phénoménologie à la croisée des arts*, ed. P. Limido-Heulot, Presses Universitaires de Rennes, AEsthetica, Rennes, 2013.
- W. Miskiewicz, *"Archives philosophique multilingues à l'époque du numérique: Le projet Archives e-LV"*. In: Patrice Bourdelais, *Institut des sciences humaines et sociales CNRS, dir. la lettre de l'INSHS*, tome 18. – La tribune d'ADONIS. – Paris: INSHS, 2012. – p. 18-20.
- W. Miskiewicz, *"Quand les technologies du Web contournent la barrière linguistique: Archives e-LV."*, Synergies Revues, vol. 1, n° 1. – Synergies Pologne n°spécial 2, 2011, p. 81-91. – ISSN: 1734-4387.
- E. Pierazzo, 'Digital genetic editions: the encoding of time in manuscript transcription'. *Text Editing, Print and the Digital World, Digital Research in the Arts and Humanities*. M. Deegan and K. Sutherland (eds.), Ashgate: Aldershot, 2008, pp. 169–186.
- E. Pierazzo, P. A. Stokes. 'Putting the text back into context: a codicological approach to manuscript transcription'. *Kodikologie und Paläographie im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*. M. Rehbein, T. Schaßan, P. Sahle (eds.) Norderstedt: Books on Demand, 2011, pp. 397-424.
- E. Pierazzo, "Digital Genetic Editions: The Encoding of Time in Manuscript Transcription." *Text Editing, Print and the Digital*

World. Ed. Marilyn Deegan and Kathryn Sutherland. Aldershot: Ashgate, 2009. 169–186.

- E. Pierazzo and M. Rehbein, *Documents and Genetic Criticism TEI Style*. TEI Consortium, 2010. [http://www.tei-c.org/SIG/Manuscripts/genetic.html].
- F. Rastier, *Arts et sciences du texte*. Paris: Presses Universitaires de France, 2001.
- A.H. Renear & D. Dubin, "Three of the four FRBR group 1 entity types are roles, not types" in Grove, A. (ed), *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, Milwaukee, WI.
- Twardowski, Kazimierz, *Actions and products. Comments on the Border Aera of Psychology, Grammar and Logic*, dans J.Pelc, *Semiotics in Poland*. 1894-1969, Dordrecht, Reidel, 1979, p. 13-27.
- TEI: Text Encoding Initiative. TEI Consortium, 2010. [http://www.tei-c.org]. Manuscript Description: [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/MS.html].
- Jean-Pierre Balpe, «ÉCRITURE», *Encyclopædia Universalis* [en ligne], consulté le 30 mars 2013. URL: http://www.universalis.fr/encyclopedie/ecriture/
- Roland Barthes, «TEXTE THÉORIE DU», *Encyclopædia Universalis* [en ligne], consulté le 30 mars 2013. URL: http://www.universalis.fr/encyclopedie/theorie-du-texte/
- *Fonctions et Produits* dans les Édition e-LV: la publication en ligne des versions polonaise, allemande et française des manuscrits de Twardowski encodées TEI. http://www.elv-akt.net/ressources/editions.php

Modelling frequency data: methodological considerations on the relationship between dictionaries and corpora

Moerth, Karlheinz; Budin, Gerhard; Romary, Laurent

The research questions addressed in our paper stem from a bundle of linguistically focused projects which –among other activities– also create glossaries and dictionaries which are intended to be usable both for human readers and particular NLP applications. The paper will comprise two parts: in the first section, the authors will give a concise overview of the projects and their goals. The second part will concentrate on encoding issues involved in the related dictionary production. Particular focus will be put on the modelling of an encoding scheme for statistical information on lexicographic data gleaned from digital corpora.

The mentioned projects are tightly interlinked, are all joint endeavours of the Austrian Academy of Sciences and the University of Vienna and conduct research in the field of variational Arabic linguistics. The first project, the Vienna Corpus of Arabic Varieties (VICAV), was already started two years ago on the basis of a low budget scheme and was intended as an attempt at setting up a comprehensive research environment for scholars pursuing comparative interests in the study of Arabic dialects. The evolving VICAV platform aims at pooling linguistic research data, various language resources such as language profiles, dictionaries, glossaries, corpora, bibliographies etc. The second project goes by the name of Linguistic Dynamics in the Greater Tunis Area: A Corpus-based Approach. This three-year project which is financed by the Austrian Science Fund aims at the creation of a corpus of spoken youth language and the compilation of a diachronic dictionary of Tunisian Arabic. The third project which has grown out of a master's thesis deals with the lexicographic analysis of the Wikipedia in Egyptian vernacular Arabic. In all these projects, digital data production relies on the Guidelines of the TEI (P5), both for the corpora and the dictionaries. The dictionaries compiled in the framework of these projects are to serve research as well as didactic purposes.

Using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. Our paper will not resume the TEI vs. LMF vs. LexML vs. Lift vs. ... discussion (cf. Budin et al. 2012) and assumes that the TEI dictionary module is sufficiently well-developed to cope with all requirements needed for the purposes of our projects. The basic schema used has been tested in several projects for various languages so far and will furnish the foundation for the intended customisations.

Lexicostatistical data and methods are used in many fields of modern linguistics, lexicography is only one of them. Modern-time dictionary production relies on corpora, and statistics–beyond any doubt–play an important role in lexicographers' decisions when selecting lemmas to be included in dictionaries, when selecting senses to be incorporated into dictionary entries and so forth. However, lexicostatistical data is not only of interest for the lexicographer, it might also be useful to the users of lexicographic resources, in particular digital lexicographic resources. The question as to how to make such information available takes us to the issue of how to encode such information.

Reflecting on the dictionary–corpus–interface and on the issue of how to bind corpus-based statistical data into the lexicographic workflow, two prototypical approaches are conceivable: either statistical information can statically be embedded in the dictionary entries or the dictionary provides links to services capable of providing the required data. One group of people working on methodologies to implement functionalities of the second type is the Federated Content Search working group, an initiative of the CLARIN infrastructure which strives to move towards enhanced search-capabilities in locally distributed data stores (Stehouwer et al. 2012). FCS is aiming at heterogeneous data, dictionaries are only one type of language resources to be taken into consideration. In view of more and more dynamic digital environments, the second approach appears to be more appealing. Practically, the digital workbench will remain in need of methods to store frequencies obtained from corpus queries, as human intervention will not be superfluous any time soon. Resolving polysemy, grouping of instances into senses remain tasks that cannot be achieved automatically.

Which parts of a dictionary entry can be considered as relevant? What is needed is a system to register quantifications of particular items represented in dictionary entries. The first thing that comes to mind are of course headwords, lemmata. However, there are other constituents of dictionary entries that might be furnished with frequency data: inflected wordforms, collocations, multi word units and particular senses are relevant items in this respect.

The encoding system should not only provide elements to encode these, but also allow to indicate the source from which the data were gleaned and how the statistical information was created. Ideally, persistent identifiers should be used to identify not only the corpora but also the services involved to create the statistical data.

We basically see three options to go about the encoding problem as such: (a) to make use of some TEI elements with very stretchable semantics such as `<note>`, `<ab>` or `<seg>` and to provide them with **`@type`** attributes, (b) to make use of TEI feature structures or (c) to develop a new customisation. We will discuss why we have discarded the first option, will present a provisional solution on the basis of feature structures and discuss pros-and-cons of this approach. As is well known, feature structures are a very versatile, sufficiently well-explored tool for formalising all kinds of linguistic phenomena. One of the advantages of the `<fs>` element is that it can be placed inside most elements used to encode dictionaries.

```
<entry xml:id="mashcal_001">
<form type="lemma">
<orth xml:lang="ar-arz-x-cairo-vicavTrans">maš#al</orth>
<orth xml:lang="ar-arz-x-cairo-arabic">مشغل</orth>
<fs type="corpFreq">
<f name="corpus" fVal="#wikiMasri"/>
<f name="frequency">
<numeric value="6"/>
</f>
</fs>
</form>
<gramGrp>
<gram type="pos">noun</gram>
<gram type="root" xml:lang="ar-arz-x-cairo-vicavTrans"
>š#l</gram>
</gramGrp>
<form type="inflected" ana="#n_pl">
<orth xml:lang="ar-arz-x-cairo-vicavTrans">mašā#il</orth>
<orth xml:lang="ar-arz-x-cairo-arabic">مشاغل</orth>
<fs type="corpFreq">
```

```
<f name="corpus" fVal="#wikiMasri"/>
<f name="frequency">
<numeric value="2"/>
</f>
</fs>
</form>
</entry>
```

The paper will be concluded by first considerations considering a more encompassing ODD based solution. We hope the work could lead to the introduction of a comprehensive set of descriptive objects (attributes and element) to describe frequencies in context, encompassing: reference corpus, size of reference corpus, extracted corpus, size of extracted corpus and various associated scores (standard deviation, t-score, etc.).

Selected references

[1] Banski, Piotr, and Beata Wójtowicz. 2009. FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries. In TEI-MM, Ann Arbor. (http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz-TEIMM-presentation.pdf)

[2] Bel, Nuria, Nicoletta Calzolari, and Monica Monachini (eds). 1995. Common Specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B. Pisa.

[3] Budin, Gerhard, Stefan Majewski, and Karlheinz Mörth. 2012. Creating Lexical Resources in TEI P5. In jTEI 3.

[4] Hass, Ulrike (ed). 2005. Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz. Berlin; New York: W. de Gruyter.

[5] Romary, Laurent, Susanne Salmon-Alt, and Gil Francopoulol. 2004. Standards going concrete : from LMF to Morphalou. In Workshop on enhancing and using electronic dictionaries. Coling 2004, Geneva.

[6] Romary, Laurent, and Werner Wegstein. 2012. Consistent Modeling of Heterogeneous Lexical Structures. In jTEI 3.

[7] Sperberg-McQueen, C.M., Lou Burnard, and Syd Bauman (eds). 2010. TEI P5: Guidelines for Electronic Text Encoding and

Interchange. Oxford, Providence, Charlotteville, Nancy. (http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf)

[8] Stehouwer, Herman, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated Search: Towards a Common Search Infrastructure. In: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Mariani, Joseph (eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). Istanbul.

[9] Werner Wegstein, Werner, Mirjam Blümm, Dietmar Seipel, and Christian Schneiker. 2009. Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch. (http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf)

A Saussurean approach to graphemes declaration in charDecl for manuscripts encoding

Monella, Paolo

The current approach of TEI to the issue of graphemes encoding consists in recommending to use the Unicode standard. This is sufficient, on the practical side, when we encode printed documents based on post-Gutenberg writing systems, whose set of graphic signs (graphemes, diacritics, punctuation etc.) can be considered standard and implicitly assumed as known.
However, each historical textual document like a medieval manuscript or an ancient inscription features a specific writing system, different from the standard emerged after the invention of print.
This implies that the TEI 'Unicode-compliance' principle is not sufficient to define graphemes in pre-print writing systems. Let us assume that manuscript A has two distinct graphems 'u' and 'v', while manuscript B has only one 'u' grapheme. If we identified both the 'u' of the first manuscript

and the 'u' of the second manuscript with the same Unicode codepoint (U+0075), our encoding would imply that they are the same grapheme, while they are not. Each of them, instead, is defined contrastively by the net of relations in the context of its own writing system, and the net of contrastive relations of manuscript A is different from that of manuscript B, as the latter does not have a 'u/v' distinction. This is even more evident with other graphic signs such as punctuation, whose expression (shape) and content (value) varied enormously through time.

This is why Tito Orlandi (2010) suggests to declare and define formally, for each document edited (e.g. a manuscript), each graphic sign that the encoder decides to distinguish, identify and encode in his or her digital edition. The natural place for this description seems to be the `charDesc` element within the TEI Header.

However, a specific technical issue arises, that I shall discuss in this paper: the TEI gaiji module only allows for a description of 'non-standard characters', i.e. graphemes and other signs not included in Unicode.

To my knowledge, there is currently no formal way in TEI to declare the specific set of 'standard' Unicode characters used in a digital edition and to define the specific value of the corresponding graphemes in the ancient document's writing system.

This is due to the current TEI general approach to the encoding of 'characters'. The TEI Guidelines currently suggest that encoders define as few 'characters' as possible, while I am suggesting that they should declare and define all encoded signs.

Possible solutions to this specific issue will be examined in this paper. I shall discuss possible changes to the TEI schema to allow for Unicode characters to be re-defined in the specific context of TEI transcriptions of ancient textual sources. Finally, I shall suggest how this might change the general approach towards the issue of graphemes encoding in the TEI Guidelines. I think that, at least in the case of the encoding of ancient documents, it should be recommended that all graphic signs identified, and not only 'non-standard' ones, be formally declared and defined.

To be more specific, the `glyph` element in the `charDesc` currently allows the encoder to freely define as many glyphs (i.e. allographs)

as desired. It is not required, however, to give a complete list of the allographs of a manuscript. The `g` elements pointing to `glyph` definitions are meant to annotate individual *tokens*/instances of a given character (i.e. grapheme) in the `body` of the transcription, but it is not possible to annotate, i.e. to *describe* that character/grapheme as a *type* in the `charDesc`, if it is encoded by means of an existing Unicode codepoint (like the very common 'u', U+0075).

The Guidelines currently recommend, instead, to define characters/graphemes in the `charDesc` section of the TEI Header by means of `char` elements *only* if they are not already present in the Unicode character set. The encoder cannot *re-define* or annotate the specific value of that character in a manuscript's graphical system if that character exists in Unicode.

This is not only a matter regarding documentation, i.e. the Guidelines' current policy on character and glyph description. Let us imagine that an encoder decided to follow the approach suggested by Orlandi and to prepend to the transcription of a manuscript a complete and formal list of *all* graphemes and/or allographs identified in the manuscript by means of `char` and/or `glyph` elements respectively. This would imply overriding even the most common Unicode characters, such as 'a', 'b' and 'c', thus overhauling the approach suggested by the Guidelines – but would still be theoretically feasible on the basis of the current gaiji module. However, if he or she decided to define every character or glyph in the `charDesc` section, they would then be required to encode *each single* grapheme or allograph in the `body` of the transcription by means of a `g` element (or by means of an XML entity expanding to that element).

In the model that I am advocating, if the editor is providing a transcription of a pre-Gutenbergian primary source the Guidelines shoud recommend to formally list and briefly describe in `charDesc` *all* characters and glyphs (i.e. graphemes and allographs) identified. The gaiji module should also provide a mechanism by which, for example:

- The encoder can decide to encode the 'u/v' grapheme of a manuscript simply by means of Unicode character U+0075 ('u');
- He or she must give a brief formal definition of the value that the grapheme encoded with Unicode codepoint U+0075 has *in the*

encoded manuscript (e.g. as *not* distinct from 'v') by means of the `char` element in `charDesc`;

- In the `body` of the transcription, they can simply transcribe that grapheme by means of Unicode character U+0075 (one keystroke).

Bibliography

- Baroni A. (2009). *La grafematica: teorie, problemi e applicazioni*, Master's thesis, Università di Padova. <http://unipd.academia.edu/AntonioBaroni/Papers/455456/La_grafematica_teorie_problemi_e_applicazioni>. [last retrived 10.03.2013].
- Mordenti R. (2001). *Informatica e critica dei testi*, Bulzoni.
- Mordenti R. (2011). *Paradosis. A proposito del testo informatico*, Accademia Nazionale dei Lincei.
- Monella P. (2012). *In the Tower of Babel: modelling primary sources of multi-testimonial textual transmissions*, a talk delivered at the London Digital Classicist Seminars 2012, Institute of Classical Studies, London, on 20.07.2012. <http://www.digitalclassicist.org/wip/wip2012.html>. [last retrieved 17.03.2013].
- Orlandi T. (1999). *Ripartiamo dai diasistemi*, in I nuovi orizzonti della filologia. Ecdotica, critica testuale, editoria scientifica e mezzi informatici elettronici, Conv. Int. 27-29 maggio 1998, Accademia Nazionale dei Lincei, pp. 87-101.
- Orlandi T. (2010). *Informatica testuale. Teoria e prassi*, Laterza.
- Perri A. (2009). *Al di là della tecnologia, la scrittura. Il caso Unicode*. «Annali dell'Università degli Studi Suor Orsola Benincasa» 2, pp. 725-748.
- Sampson G. (1990). *Writing Systems: A Linguistic Introduction*, Stanford University Press.
- Wittern C. (2006). *Writing Systems and Character Representation*, in L. Burnard, K. O'Brien O'Keeffe, J. Unsworth, edd., Electronic Textual Editing, Modern Language Association of America.

Texts and Documents: new challenges for TEI interchange and the possibilities for participatory archives

Muñoz, Trevor; Viglianti, Raffaele; Fraistat, Neil

Abstract

The introduction in 2011, of additional "document-focused" (as opposed to "text-focused") elements represents a significant additional commitment to modeling two distinct ontologies within the Text Encoding Initiative (TEI) Guidelines, and places increased strain on the notion of "interchange" between and among TEI data modeled according to these two approaches. This paper will describe challenges encountered by members of the development and editorial teams of the *Shelley-Godwin Archive* (S-GA) in attempting to produce TEI-encoded data reflecting both "document-focused" and "text-focused" approaches through automated conversion. S-GA started out, like most electronic literary archives, with the primary goal of providing users access to rare and widely dispersed primary materials, but increasingly the direction of the project will be to take advantage of the tremendous potential of its multi-layered architecture to re-conceptualize and design the whole as a work-site, or what some are calling an "animated archive," whose ultimate goal is to make the S-GA material massively addressable in a form that encourages user curation and exploration. The ability to convert from "document-focused" to "text-focused" data—from work-site to publication—will partly determine how participatory the archive can be.

Background & Motivation

The *Shelley-Godwin Archive* is a project involving the Maryland Institute for Technology in the Humanities (MITH) and the Bodleian, British, Huntington, Houghton, and New York Public libraries that will contain the works and all known manuscripts of Mary Wollstonecraft, William Godwin, Percy Bysshe Shelley, and Mary Wollstonecraft Shelley. We wish to produce two distinct representations of the S-GA materials so as (1) to provide rigorous, semi-diplomatic transcriptions of the fragile

manuscripts for those with an interest in the compositional practices of what has been called "England's First Family of Literature" and (2) to make available clear "reading texts" for those who are primarily interested in the final state of each manuscript page.

The start of text encoding work on the S-GA coincided with the addition of new "document-focused" elements to the TEI in the release of P5 version 2.0.1. Given that the majority of materials in the collection consist of autograph manuscripts, the project team quickly adopted several of these new elements into its TEI customization. The "genetic editing" approach has served the project well—allowing the encoding scheme to target features of the documents that are of greatest interest to the scholarly editors and to rigorously describe often complicated sets of additions, deletions, and emendations that will support further scholarship on the composition process of important literary works. The work of automating the production of usable "reading texts" encoded in "text-focused" TEI markup from data that is modeled according to a "document-focused" approach has proven much more challenging.

Encoding Challenges

The conflict between representing multiple hierarchies of content objects and the affordances of XML is well known and the TEI Guidelines discuss several possible solutions. One of these solutions is to designate a primary hierarchy and to represent additional hierarchies with empty milestone elements that can be used by some processing software to "reconstruct" an alternate representation of the textual object. The approach taken by the S-GA team to produce both "document-focused" and "text-focused" TEI data is a version of the milestone-based approach. The document-focused, "genetic editing" elements form the principal hierarchy (consisting of "<surface>," "<zone>," "<line>," etc.) and milestone elements are supplied to support automatic conversion to "text-focused" markup (which will contain elements such as " ," "<p>," "<lineGrp>," etc.).

This solution places increased burden on document encoders to maintain "correctness," thus potentially lowering data consistency and quality. For instance, empty element milestones representing the beginning and ending of textual features have no formal linkages as part of the document

tree. Encoders must supply identifiers and pointers to indicate these linkages. Validating that these identifiers and pointers pair correctly must be accomplished with some mechanism other than the RelaxNG validation that verifies most other elements of the document structure. As noted above, managing multiple hierarchies through the use of milestones is not new. We do argue that the introduction of additional "document-focused" elements in the TEI increases the scope for projects to produce data that reflect two divergent ontologies and thus to encounter the difficulties involved in this "workaround."

More importantly, the use of the milestone strategy decreases the reusability of the data. For example, to support automated conversion from "document-focused" to "text-focused" data representations, the S-GA team needed to go beyond purpose-built milestone elements like "<delSpan>" and "<addSpan>" and, in effect, semantically overload the general purpose "<milestone>" element. The value of an attribute on "<milestone>" indicates which "text-focused" element is intended to appear in a particular location. This solution is explained in the documentation and the convention used would be (we think) evident after cursory examination. Nonetheless, we are forced to add markup to the "document-focused" data which makes it more unique to the S-GA project and less easily consumable by future users with different goals. This is even more troubling because the "document-focused" data is the true work-site where we hope to invite future collaborators to engage and extend the project.

Maintainability & Provenance Challenges

To avoid the conceptual and technical challenges involved in automating the transformation between "text-focused" and "document-focused" representations, the two sets of data could have each been created by hand and maintained separately. Indeed, this is the approach followed by the *Digitale Faustedition* project, where a distinction between what the project calls "documentary" and "textual" transcription was considered necessary not only as a reaction to encoding problems, but also as a practical application of theoretical distinctions between documentary record and editorial interpretation. The *Faustedition* project team, however, still encountered technical challenges when trying to

correlate and align these two transcriptions automatically. Use of collation and natural language processing tools helped with this problem, but eventually more manual intervention was needed (Brüning *et al*. 2013).

The S-GA team felt that maintaining two data sets representing different aspects of the textual objects would have led to serious data consistency, provenance, and curation problems. As the example of the *Faustedition* project shows, separate representations must be kept in sync with project-specific workflows developed for this purpose. In the case of S-GA, documentary transcription is the main focus; the greatly increased cost and time involved in also maintaining a textual transcription would have reduced the size of the corpus that could be encoded and thus the amount of materials from the archive that could be made fully available under the current phase of the project.

Presentation Challenges

The display and presentation of "document-focused" encoding is another technical challenge introduced by the new TEI elements; to provide a diplomatic transcription, a TEI-to-HTML transformation is not trivial—often times limited by HTML's own capabilities. A canvas-based system, such as PDF or SVG, is better suited for presenting document-focused encoding.

S-GA is developing and using a viewer for SharedCanvas, a technology developed at Stanford University, that allows editors (and potentially future users) to construct views out of linked data annotations. Such annotations, expressed in the Open Annotation format, relate images, text, and other resources to an abstract "canvas". In S-GA, "document-focused" TEI elements are mapped as annotations to a SharedCanvas manifest and displayed. Further layers of annotations can be added dynamically, for example search result highlights as well as user comments and annotations. The engagement of students and other scholars will be driven by the possibility of creating annotations in the Open Annotation format, so that any SharedCanvas viewer will be able to render them. It remains a matter for the future development of the project to understand whether some annotations can be added dynamically to the source TEI, especially those pertaining transcription and editorial statements.

Consequences

The attempt to automatically generate "text-focused" markup from "document-focused" markup forced the project team to confront the intellectual challenges which the introduction of the genetic editing element set makes urgent. The larger stakes involved were made clear to the project team during our recent experiments with the distributed TEI encoding of the manuscripts of *Frankenstein* and *Prometheus Unbound* by graduate students at the University of Maryland and the University of Virginia. The attempt to bring additional encoders of various skill levels into the editing and encoding of the Shelley-Godwin materials revealed the importance of being able to convert from "document-focused" to "text-focused" data because this ability will partly determine how participatory the archive can be. The Digital Humanities is now undergoing what might be called a "participatory turn" that poses for the creators of digital literary archives such questions as (1) How can humanists best curate and explore our datasets? (2) How can we bring our research into the graduate and undergraduate classroom, including the process of text encoding?; and (3) How can we fruitfully engage the public, "citizen humanists," in the work of the humanities? The potential to address these larger questions will necessarily proceed from the way in which the TEI community grapples with the modeling challenges of supporting two distinct ontologies of textual objects.

Acknowledgements

The *Shelley-Godwin Archive* is a collaborative endeavor. In developing the ideas in this paper, we have benefited from discussions with Travis Brown, Jim Smith, David Brookshire, Jennifer Guiliano, and other members of the Shelley-Godwin Archive project team.

Bibliography

- Bauman, S. "Interchange Vs. Interoperability." Montréal, QC. Accessed April 7, 2013. doi:10.4242/BalisageVol7.Bauman01.
- Brüning, G., et al. Multiple Encoding in Genetic Editions: The Case of "Faust" http://jtei.revues.org/697
- Pierazzo, E. A rationale of digital documentary editions http://llc.oxfordjournals.org/content/26/4/463

• Sanderson, R., et al. SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination http://arxiv.org/pdf/1104.2925v1.pdf

Beyond nodes and branches: scripting with TXSTEP

Ott, Wilhelm; Ott, Tobias

Two years ago, at the 2011 TEI members meeting in Würzburg, we presented a first feasibility study and preliminary model of TXSTEP, an open source, XML based scripting language which will make available the power of TUSTEP by an up-to-date and self-explaining user interface. TUSTEP itself is known as a very powerful set of tools for the processing, analysis and publication of texts, meeting the requirements of scholarly research - and at the same time as having a very steep learning curve, an unfamiliar command-line based user interface and a documentation which is avaliable in German only.
TXSTEP breaks down these barriers to the usability of these tools. It makes available them to the growing e-humanities community, offering them a powerful tool for tasks which can not easily be performed by the scripting tools commonly used for this purpose. At the same time, it allows to integrate the mentioned tools into existing XSL-based workflows.
Compared to the original TUSTEP command language, TXSTEP

- offers an up-to-date and established syntax
- allows to draft respective scripts using the same XML-editor as when writing XSLT or other XML based scripts
- lets you enjoy the typical benefits of working with an XML editor, like content completion, highlighting, showing annotations, and, of course, verifying your code,
- offers - to a certain degree - a self teaching environment by commenting on the scope of every step.

TXSTEP has in the meantime been subjected to a closer examination by Michael Sperberg-McQueen regarding its overall goal and design, the syntax and structure of the XML command language, including details of naming and style, operating system dependencies, and its positioning within the XML software ecosystem. His critics and proposals - and his very encouraging final remarks - have been very helpful for the further work on the system in the past two years. As a result, we can now present and offer for download a running system containing the modules described below in the current version 0.9.

In the February 2012 issue of the TEI Journal, Marjorie Burghart and Malte Rehbein reported on the results of a survey they had carried out and which "highlight the need for user-friendly, bespoke tools facilitating the processing, analysis, and publishing of TEI-encoded texts".

With this paper, we want to show how TXSTEP, though not restricted to work with TEI- or XML-encoded texts, could meet a great deal of the mentioned needs of text based research.

The term "user-friendly", used in the report, suggests that a typical user will be guided by an intuitive interface to ready-made solutions for a problem foreseen by the developer of the respective tool. "But", to quote Martin Müller from Northwestern University, "that is not what happens in research".

TXSTEP aims at being "user-friendly" above all to the "exploratory" user who is seriously engaged in research. The tools he needs are different: of course, they have to avoid the need of elementary programming. TXSTEP therefore offers program modules for the very basic or elementary operations of text data handling. These modules allow for further adaptation, (e.g., for defining the collating sequence required for sorting the words for non-English texts). It is possible to run each of these modules separately, but also to team them with any other module of the system.

The TXSTEP modules include:

- collation of different versions of a text, the results being stored (including TEI-based tagging) in a file for further automatic processing, in addition to being available for eye inspection;
- text correction and enhancement not only by an interactive editor, but also in batch mode, e.g. by means of correction instructions

prepared beforehand (by manual transcription, or by program, e.g. the collation module);
- decomposing texts into elements (e.g. word forms) according to rules provided by the user, preparing them for sorting according to user-defined alphabetical rules and other sorting criteria);
- building logical enities (e.g. bibliographic records) consisting of more than one element or line of text and preparing them for sorting;
- sorting such elements or entities;
- preparing indexes by generating entries from the sorted elements;
- transforming textual data by selecting records or elements, by replacing strings or text parts, by rearranging, complementing or abbreviating text parts;
- integrating additional information into a file by means of acronyms;
- updating crossreferences;
- (by including respective native TUSTEP scripts:) professional typesetting, meeting ambitious layout demands as needed for critical editions.

As the output of any one of these modules may serve as input to any other module (including XSLT-stylesheets), the range of research problems for which this system may be helpful is quite wide.
A set of modules like these is rather not appropriate for the occasional end user; its purpose is to make the professional user or the serious humanities scholar independent of foreign programming, even for work not explicitely foreseen by the developers, and to give him at the same time complete control over every detail of the data processing part of his project. It is the user himself who, instead of using a black box, defines in every detail the single steps to be performed.
It is obvious that the use of a modular system like this differs essentially from the use of tools that claim intuitive usability. It differs in two points:

- First, it requires previous learning, and
- Second, it requires to analyze a problem before starting to solve it.

It shares these features with other scripting languages.

While there is usually no way for escaping the second point, TXSTEP offers a remedy for the first problem.
How "user-friendly" this can be for professional use in a research environment, we will demonstrate live by means of some elementary examples of text handling and text analysis which can not easily be solved with existing XML tools.

Bibliography

- Eberhard Karls Universität Tübingen, Zentrum für Datenverarbeitung: TUSTEP. Tübinger System von Textverarbeitungsprogrammen. Version 2013. Handbuch und Referenz. http://www.tustep.uni-tuebingen.de/pdf/handbuch.pdf
- Tübinger System von Textverarbeitungs-Programmen TUSTEP. http://www.tustep.uni-tuebingen.de
- TXSTEP - an integrated XML-based scripting language for scholarly text data processing. In: digital humanities 2012. Conference Abstracts.
- Creating, enhancing and analyzing TEI files: the new, XML-based version of TUSTEP. In: Philology in the Digital Age. Annual TEI Conference, Würzburg 2011.
- XSTEP - die XML-Version von TUSTEP. http://www.xstep.org

TEI in LMNL: Implications for modeling

Piez, Wendell

What might TEI look like if it were not based in XML? This is not simply an aesthetic question (TEI using a different sort of tagging syntax) but a very practical one, inasmuch as XML comes with limitations and encumbrances along with its strengths. Primary among these (as has been recognized since the first applications of SGML to text encoding

in the humanities) is the monolithic hierarchy imposed by the XML data model. Texts of interest to the humanistic scholar frequently have multiple concurrent hierarchies (in addition to the 'logical' structure of a text generally presented in XML, we have physical page structures; dialogic and narrative structures; the grammar of natural language; rhetorical and verse structures; etc. etc.), as well as 'arbitrary overlap' — constructs found in the text stream that form no hierarchy at all, such as ranges to be indexed or annotated, which can overlap freely both with other structures and with one another.

Of course, TEI proposes mechanisms for dealing with these (in an entire chapter of the Guidelines devoted to this topic), and since the introduction of XPath/XSLT 2.0 along with XQuery, we have more capable means for processing them. But the code we have to write is complex and difficult to develop and maintain. What if we didn't have to work around these problems?

LMNL) offers such a model, and a prototype LMNL processing pipeline — Luminescent, supporting native LMNL markup on an XML/XSLT platform — offers a way to explore these opportunities. TEI XML documents can be processed programmatically to create LMNL markup, with its representations of overlap (whether using milestones, segmentation, or standoff) converted into direct markup representations. Once in LMNL syntax, ranges and annotation structures can be used to refactor complex XML structures into simpler forms directly correspondent (i.e., without the overhead of pointers) to the textual phenomena they apply to. In particular, the LMNL model has two features that (separately and together) enable significant restructuring and resolution of modeling issues, exposing complexities as they are rather than hiding phenomena (which in themselves may be simple or complex) behind necessary complexities of syntax:

- Because ranges can overlap freely, families of related ranges emerge, each family overlapping others, but no ranges within a single family overlapping other ranges in the same family. (And here we have *multiple concurrent hierarchies*, although in LMNL the hierarchical relation among ranges in a single family is only implicit.) For example, one set of ranges represents a

clean logical hierarchy of books, chapters, sections and paragraphs, while another represents the pagination of a physical edition, while a third represents a narrative structure. LMNL processing can disentangle these from one another, rendering any of them as a primary ('sacred') hierarchy in an XML version.

By the same token, it becomes possible to discern (through analysis of which ranges overlap others of the same or different types) where overlap is truly arbitrary: where, that is, the information indicated by a range (such as an annotated or indexed span) must be permitted to overlap others even of the same type. In other words, typologies of ranges and range types emerge, that both relate them systematically to one another, or deliberately permit them to be unrelated.

- Since LMNL annotations can be structured and their contents marked up, annotations can take on more of the burden of data capture than is easily or gracefully done with XML attributes. It becomes possible once again, even at significant levels of complexity, to make a broad distinction between the text being marked up, and the apparatus attached to the text.

Demonstrations will be offered, showing both TEI data in LMNL, and the kinds of outputs (in plain text, HTML, SVG or XML including TEI) that can be generated from it.

Bibliography

This is only a partial (in fact quite incomplete) bibliography of work in this area.

- David Barnard, Ron Hayter, Maria Karababa, George Logan and John McFadden. 1988. *SGML-Based Markup for Literary Texts: Two Problems and Some Solutions*. Computers and the Humanities, Vol. 22, No. 4 (1988), pp. 265-276.
- David Barnard, Lou Burnard, Jean-Pierre Gaspart, Lynne A. Price, C. M. Sperberg-McQueen and Giovanni Battista Varile. 1995. *Hierarchical Encoding of Text: Technical Problems and SGML Solutions*. Computers and the Humanities, Vol. 29, No. 3, The Text Encoding Initiative: Background and Context (1995), pp. 211-231.

- CATMA: Computer Aided Textual Markup and Analysis. See http://www.catma.de/.
- James H. Coombs, Allen H. Renear, and Steven J. DeRose. 1987. *Markup Systems and The Future of Scholarly Text Processing*. Communications of the ACM, 30:11 933-947 (1987).
- Claus Huitfeldt. 1994. *Multi-Dimensional Texts in a One-Dimensional Medium*. Computers and the Humanities, Vol. 28, No. 4/5, Humanities Computing in Norway (1994/1995), pp. 235-241.
- Paolo Marinelli, Fabio Vitali, and Stefano Zacchiroli. 2008. *Towards the unification of formats for overlapping markup*. At http://upsilon.cc/~zack/research/publications/nrhm-overlapping-conversions.pdf.
- Wendell Piez. 2004. *Half-steps toward LMNL*. In Proceedings of Extreme Markup Languages 2004. See http://conferences.idealliance.org/extreme/html/2004/Piez01/EML2004Piez01.html.
- Wendell Piez. 2008. *LMNL in Miniature: An introduction*. Amsterdam Goddag Workshop, December 2008. Presentation slides at http://piez.org/wendell/LMNL/Amsterdam2008/presentation-slides.html.
- Wendell Piez. 2010. *Towards Hermeneutic Markup: an Architectural Outline*. Presented at Digital Humanities 2010 (King's College, London), July 2010. Abstract and slides at http://piez.org/wendell/dh2010/index.html.
- Wendell Piez. 2011. *TEI Overlap Demonstration*. At http://piez.org/wendell/projects/Interedition2011/.
- Wendell Piez. 2012. *Luminescent: parsing LMNL by XSLT upconversion*. Presented at Balisage: The Markup Conference 2012 (Montréal, Canada), August 2012. In Proceedings of Balisage: The Markup Conference 2012. Balisage Series on Markup Technologies, vol. 8 (2012). doi:10.4242/BalisageVol8.Piez01.
- Allen Renear, Elli Mylonas and David Durand. 1993. *Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies*. At http://www.stg.brown.edu/resources/stg/monographs/ohco.html.

- Desmond Schmidt. 2010. *The inadequacy of embedded markup for cultural heritage texts*. Literary and Linguistic Computing (2010) 25(3): 337-356. doi: 10.1093/llc/fqq007.
- C. M. Sperberg-McQueen. 1991. *Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts*. Literary and Linguistic Computing, Vol. 6, No 1, 1991.
- C. M. Sperberg-McQueen. 2006. *Rabbit/duck grammars: a validation method for overlapping structures*. In Proceedings of Extreme Markup Languages 2006, Montreal, August 2006. At http://www.idealliance.org/ papers/extreme/proceedings/html/2006/SperbergMcQueen01/EML2006SperbergMcQueen01.html.
- M. Stührenberg and D. Goecke. 2008. *SGF - An integrated model for multiple annotations and its application in a linguistic domain*. Presented at Balisage: The Markup Conference 2008 (Montréal, Canada), August 2008. In Proceedings of Balisage: The Markup Conference 2008. Balisage Series on Markup Technologies, vol. 1 (2008). doi: 10.4242/BalisageVol1.Stuehrenberg01.
- M. Stührenberg and D. Jettka. 2009. *A toolkit for multi-dimensional markup - The development of SGF to XStandoff*. Presented at Balisage: The Markup Conference 2009 (Montréal, Canada), August 2009. In Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009). doi: 10.4242/BalisageVol3.Stuhrenberg01.
- Jeni Tennison and Wendell Piez. 2002. *The Layered Markup and Annotation Language (LMNL)*. Extreme Markup Languages 2002.
- Jeni Tennison. 2007. *Creole: Validating Overlapping Markup*. Presented at XTech 2007. http://assets.expectnation.com/15/event/1/Creole_%20Validating%20Overlapping%20Markup%20_Prince%20PDF%20version_.pdf

  Text Encoding Initiative (TEI). *P5: Guidelines for Electronic Text Encoding and Interchange, chapter 20, Non-hierarchical Structures*. At http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html.

  XStandoff. http://www.xstandoff.net/.

TEI at Thirty Frames Per Second: Animating Textual Data from TEI Documents using XSLT and SVG

Pytlik Zillig, Brian L.; Barney, Brett

The growing abundance of TEI-encoded texts—including some rather large-scale collections such as those associated with the Brown University Women Writers Project, Perseus Digital Library, Wright American Fiction, and the University of Michigan's Text Creation Partnership—in conjunction with an expanding palette of visualization tools, has made it possible to create graphic representations of large-scale phenomena. Visual representations, traditional examples of which include graphs, lists, concordances, tables, and charts, have often been used to bring focus to aspects that might otherwise be overlooked. That is, they are in part tools for noticing, assisting the user/reader in seeing what may be difficult or impossible to perceive in the textual flow when it is presented in the conventional manner. As Tanya Clement has recently observed, "Sometimes the view facilitated by digital tools generates the same data human beings . . . could generate by hand, but more quickly," and sometimes "these vantage points are remarkably different . . . and provide us with a new perspective on texts." And as Dana Solomon has written, "[d]ue in large part to its often powerful and aesthetically pleasing visual impact, relatively quick learning curve … and overall 'cool,' the practice of visualizing textual data has been widely adopted by the digital humanities." When used for large textual corpora, visualizations can, among numerous other possibilities, represent change over time, group common characteristics among texts, or highlight differences among them, correlated by such factors as author, gender, period, or genre. At the University of Nebraska–Lincoln's Center for Digital Research in the Humanities we have been experimenting with a new way of visualizing phenomena in TEI corpora and have created an experimental XSLT-based tool that queries TEI files and generates animated videos of the results. Using XPath and XQuery techniques, this tool makes it possible to ask specific or general questions of a corpus such as: "What is the most frequently-occurring 3-gram in each text in this writer's oeuvre?" or

"When did the poet begin to favor use of the word 'debris'?" The data are then output as scalable vector graphic (SVG) files that are converted to raster images and rendered in video at 30 frames per second. Our present goal is to test this alpha version with the writings of Walt Whitman, or, more specifically, with a particular Whitman poem.

The Whitman Archive has been producing TEI-encoded texts of Whitman's work since 2000 and offers access to a huge variety of textual data both by and about Whitman. Among these is a poor-quality 40-second recording of someone, possibly Whitman himself, reading the first four lines of one of his lesser-known poems. Even though the Archive makes it clear that the voice may not even be Whitman's, this sound recording of "Whitman" reading "America" has been surprisingly popular and compelling. It is one of the most frequently requested pages on the site and was recently the focus of an article in Slate. One reason for the recording's popularity, surely, is its immediacy; it brings Whitman's words to life, performing them in a way that they are not when users encounter the words as fixed characters on a page or screen. The sound recording also reminds us of the importance of the performative aspect of Whitman's poetry specifically and of poetry generally. Early in his career, Whitman often recited from Shakespeare and other poets for the entertainment of ferry passengers and omnibus drivers, and his lecture notes from the 1880s demonstrate that he enjoyed performing a variety of poems—both his and others'.

The visualization tool that we have developed is, at this stage, utterly experimental; we make no claims about its superiority relative to other tools or even about its worth for literary analysis. Instead, we see its value as, first, an exploration of techniques for combining TEI and SVG data into ambitious vector-based animations and, second, as a demonstration of the potential for engaging the multi-sensory and multimodal aspects of texts. "Engagement" write Fernanda Viegas and Martin Wattenberg, "—grabbing and keeping the attention of a viewer—is the key to [data visualization's] broader success." In representing the literary work as an absorbing performance, one that comprises both "data" and "art," the tool we are developing is calculated to provoke responses in both informational and aesthetic registers. Performance and provocation are perhaps not the

most efficient means of adducing, synthesizing, or rendering evidence, but they might well supplement other techniques in conveying some of the complex ways in which literary texts work.

Bibliography

- Clement, T. "Text Analysis, Data Mining, and Visualizations in Literary Scholarship" in Literary Studies in the Digital Age: An Evolving Anthology (eds., Kenneth M. Price, Ray Siemens), 2013. Modern Language Association.
- Solomon, D. "Building the Infrastructural Layer: Reading Data Visualization in the Digital Humanities." MLA 2013 Conference Presentation. url: http://danaryansolomon.wordpress.com/2013/01/08/mla-2013-conference-presentation-from-sunday-162013/
- Viegas, Fernanda, and Martin Wattenberg. "How to Make Data Look Sexy." CNN Opinion, 19 April 2011. http://www.cnn.com/2011/OPINION/04/19/sexy.

Analysis of isotopy: a hermeneutic model

Scacchi, Alessia

The presentation illustrates the analysis of isotopes in twentieth-century literature as a template of deep interpretation of texts, which increases the traditional analytical procedures, proposing an evolution of practices.

The topic fits in the broad debate involving the critical literature in the age of (re)producibility (Riva, 2011), and suggests a rethinking of models and methods in textual hermeneutics, using a digital way (Ciotti, Crupi, 2012). The novelty consists in doing narratological analysis observing its macrostructural and microstructural results (styles, lexemes, isotopes), proposing a hermeneutic template that allows semantic indexing of

families and isotopes, deductible through broad concepts: place, space, character and identity.

With the proliferation of tools and technologies that enable the increasing of text data and electronic editions in different formats (rtf, pdf, epub, oeb), decreases the hermeneutic potential triggered by computer when the text is divided into atoms of meaning (Trevisan, 2008). Besides, textual criticism is often lacking historical dimension that communicative act in literary work testifies. So the paper propose a solution to storage problems, distribution, and analysis of literary works in historical perspective, using TEI to codify some semantic features in modern texts.

The analytic practice, promoted by Crilet Laboratory at Faculty of Arts in University "Sapienza" of Rome, is aimed at expanding interpretation purpose of documents (Mordenti, 2007), with digital transcription and its redrafting in semantic markup. So, using literary and hermeneutic tags rather than philological, it develops a pragmatic combination of history and semiotics so that the digital document represents, inside, the interpretative model.

Infact, it's possible to span narrative corpus in many areas of meaning and analyse: vertically, studying the lexical sorting from maximum frequency to hapax; semantically, studying frequency and position of selected isotopes in text (Greimas, 1970); alphabetically, generating an alphabetical order to identify families of meanings. At this point, having built a system centered on text it's useful to start critical thinking adding XML markup for links to websites with historical references, within the model proposed by TEI.

Some examples of twentieth-century analytic papers are available because of my decades work in University "Sapienza" of Rome. However, considering my work as reading of a "lector in fabula" # educated and skilled # it make me able to establishe with narrative material a close relationship that also involves the author as a creator. This two characters are bound by the joint effort of giving a real and imaginary birth to the object of art. Thus, the markup should take care of an object that express meaning on two floors: reality and imagination. New technologies are helpful in the breakdown of the two levels, because of native digital architecture. In this way, the option humanities computing is emerging

as a choice of epistemology, rather than an instrumental change. A radically rethinking of concept of text appears as a new light, not a deformity designed by artificial systems, but a strict vitality, given from the automation process (Mordenti, 2007).
The paper, therefore, wants to underline the potential of textual analysis using TEI markup, providing for electronic text processing and following Segre's ideas (Ciotti, 2007; Orlandi, 2010; Fiorentino, 2011; Riva, 2011). The system built in that way would encourage the study of narrative in his historical aspects, social and cultural development, also it can be a valid tool for interpretation of textual themes and motifs related to historical context, especially in secondary schools and universities, as easy for digital born students.
Therefore, research project converge skills of a different type, related to scientific fields and disciplines of various kinds, to highlight a clear interdisciplinary nature.
To historical and literary capabilities are associated, necessary, skills of humanities computing, digital cultures (Ciotti, 2012) and textual theory, which give greater depth to the proposed analytical practice.

Bibliography

- Burnard, *Il manuale TEI Lite: introduzione alla codifica elettronica dei testi letterari*, a cura di Fabio Ciotti, Milano, Sylvestre Bonnard, 2005
- Ciotti, *Il testo e l'automa. Saggi di teoria e critica computazionale dei testi letterari*, Roma, Aracne, 2007
- Ciotti, Crupi (a cura di), *Dall'Informatica umanistica alle culture digitali. Atti del Convegno di studi in memoria di Giuseppe Gigliozzi (Roma, 27-28 ottobre 2011)*, Roma, Università La Sapienza, 2012
- Fiormonte, *Scrittura e filologia nell'era digitale*, Milano, Bollati Boringhieri, 2003
- Fiormonte, Numerico, Tomasi (a cura di), *L'umanista digitale*, Bologna, Il Mulino, 2010
- Gigliozzi, *Il testo e il computer. Manuale di informatica per gli studi letterari*, Milano, Bruno Mondadori, 1997
- Greimas, *Del senso*, Milano, Bompiani, 1970

- Holister, *Pensare per modelli*, Milano, Adelphi, 1985
- Landow, *L'ipertesto. Tecnologie digitali e critica letteraria*, trad. it. a cura di Paolo Ferri, Milano, Bruno Mondadori, 1998
- Luperini, *Il dialogo e il conflitto. Per un'ermeneutica materialistica*, Bari, Laterza, 1999
- Meyrowitz, *Oltre il senso del luogo*, Bologna, Baskerville, 1993
- Mordenti, *L'altra critica. La nuova critica della letteratura fra studi culturali, didattica e informatica*, Roma, Meltemi, 2007
- Orlandi, *Informatica testuale. Teoria e prassi*, Bari, Laterza, 2010
- Pierazzo, *La codifica dei testi*, Roma, Carocci, 2005
- Riva, *Il futuro della letteratura. Lopera letteraria nell'epoca della sua (ri)producibilità digitale*, Scriptaweb, 2011
- Szondi, *Introduzione all'ermeneutica letteraria* (1975), trad. di Bianca Cetti Marinoni, introd. di Giorgio Cusatelli, Torino, Einaudi, 1992

TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments

Silva, António Rito; Portela, Manuel

Context

Fernando Pessoa's *Book of Disquiet* (*Livro do Desassossego – LdoD*) is an unfinished book project. Pessoa wrote more than five hundred texts meant for this work between 1913 and 1935, the year of his death. The first edition of this book was published only in 1982, and another three major versions have been published since then (1990, 1998, 2010). As it exists today, *LdoD* may be characterized as (1) a set of autograph (manuscript and typescript) fragments, (2) mostly unpublished at the time of Pessoa's death, which have been (3) transcribed, selected, and organized into four different editions, implying (4) various interpretations

of what constitutes this book. Editions show four major types of variation: variation in readings of particular passages, in selection of fragments, in their ordering, and also in heteronym attribution.

Goals

The goal of the *LdoD* Archive[14] is twofold: on the one hand, we want to provide a "standard" archive where experts can study and compare *LdoD*'s authorial witnesses and their different editions; on the other hand, we want to design a virtual archive that allows both experts and non-experts to experiment with the production of different editions of *LdoD*, and also the writing of their own fragments based on *LdoD*'s original fragments.[15] Therefore, this latter goal, which is built on top of the archival goal, extends a scholarly understanding of *LdoD* as both authorial project and editorial construct to a new perspective of *LdoD* as an individual and/or community editing and writing exploratory environment based on the authorial and editorial witnesses.

Problem

Given the above set of goals, the *LdoD* Archive has to accommodate scholarly standards and requirements on digital archives, for instance the use of TEI as a specification to encode literary texts, and the virtual communities and social software features to support the social edition of *LdoD* by both other experts and non-experts.[16] This second aspect

[14] "No Problem Has a Solution: A Digital Archive of the Book of Disquiet", research project of the Centre for Portuguese Literature at the University of Coimbra, funded by FCT (Foundation for Science and Technology). Principal investigator: Manuel Portela. Reference: PTDC/CLE-LLI/118713/2010. Co-funded by FEDER (European Regional Development Fund), through Axis 1 of the Operational Competitiveness Program (POFC) of the National Strategic Framework (QREN). COMPETE: FCOMP-01-0124-FEDER-019715.

[15] A second goal of the project is to investigate the relation between writing processes and material and conceptual notions of the book. The rationale for allowing non-experts to experiment with reediting and rewriting this work originates in this second goal, and in the want to explore the collaborative dimension of the web as a reading and writing space in the context of a digital archive in ways that enhance its pedagogical, ludic, and expressive uses.

[16] The *LdoD* Archive will consider two groups of end-users and will provide tools and resources that enable engagement at different levels of complexity, from beginner to expert. Groups of beta users for the virtual editing and virtual writing features have already been

increases the need for a dynamic archive where both types of end-users can edit their own versions of *LdoD*, and write extensions of the original fragments, while the archive's experts' interpretations and analyses of *LdoD* are kept "unchanged" and clearly separated from the socialized editions and writings. In addition, it is necessary to define how the specifics of *LdoD* are represented in TEI, for instance, how do we distinguish authorial witnesses (textual records) from the editions and their respective interpretations, as when an editor assigns a fragment to heteronym Vicente Guedes while another editor assigns it to Bernardo Soares.

Solution

The solution we propose for the identified challenges is based on a TEI template to encode all authorial and editorial witnesses in TEI, and a software architecture that accommodates the traditional query and search of a digital humanities archive with functionalities of a Web2.0 approach.

Representation in TEI

We have encoded *LdoD* as a TEI Corpus containing a TEI Header for each one of the fragments. Besides the project information that is represented in the TEI Corpus, we have described properties common to the whole *LdoD*, which include (1) the set of editions; and (2) Pessoa's heteronyms. For each fragment we have encoded in a fragment header as witnesses both the original authorial sources and the four editorial sources. This approach allows us to associate interpretation metadata in the context of each witness. Users will be able to compare digital facsimile representations of authorial witnesses (and topographic transcriptions of those witnesses) to editorial witnesses. The latter can also be compared against each other in order to highlight their interpretations of the source. However, there are still some open issues. The first one is that the separation between editorial sources and authorial sources is by convention, and it is not clear how, in terms of interoperability, an external application can "understand" and process this distinction. The second aspect is related to the dynamic evolution of the archive in terms of Web2.0 requirements: how can TEI code be changed as a result of end-users' interactions with the archive?

canvassed in secondary schools and universities. These communities will allow us to better assess particular needs and define interface structure and access to contents accordingly.

Note that the traditional approach to encoding in TEI is done statically, through tools like oXygen. However, due to our requirements we want to support the evolution of *LdoD* as a continuously reeditable and rewrittable book. This means that it is necessary that we enable the addition of new virtual editions and heteronyms in the Corpus and the addition of new fragments that extend the original ones. Additionally, end-users can define their own interpretation of any of *LdoD*'s fragments, e.g. by using tags, which results in the generation of new editions of the book through the execution of a categorization algorithm. This open issue is partially addressed by the software architecture we propose in the next section.

Architecture Proposal

Most digital scholarly archives are static. By static we mean that the construction of the archive is separated from its use. The former is done using TEI and XML editors, and the latter is supported by XSLT transformations. This software architectural approach is not feasible if we want to provide Web2.0 functionality to the archive. However, we do not want to disregard what is already done in terms of encoding in TEI for the experts. Therefore the architecture needs to support the traditional encoding in TEI by the experts while enabling dynamic end-users' interactions with the platform.

The key point of the proposal is the use of an object domain model to represent the *LdoD* archive. Using this approach we, at first, transform *LdoD* encoded in TEI to the object model, and allow the visualisation and edition of this object model through a web user interface. Additionally, TEI files can be regenerated from the object model. This approach has several advantages: (1) the archives' experts continue using editor tools like oXygen to do their work; (2) end-users (experts and non-experts) can create their virtual editions and fragment extensions through the web user interface; (3) the object model preserves a semantically consistent *LdoD* archive by checking the consistency of end-users' operations; (4) interoperability can be supported by exporting the regenerated TEI files; (5) it is possible to regenerate TEI files according to different formats, for instance, it is possible to use different methods to link critical apparatus to the text.

Our proposal explores current approaches to editing in electronic environments and attempts to integrate them with TEI conceptual and processing models. The object representation of transcriptions is related with the work on data structure for representing multi-version objects (Schmidt and Colomb, 2009). We emphasize the need to have a clear separation between content and presentation in order to simplify and empower presentation tools as claimed in Schlitz and Bodine (2009). With regard to a Web2.0 for digital humanities we are indebted to proposals on cooperative annotations by Tummarello et al. (2005) and the advantages and vision of Web2.0 and collaboration in Benel and Lejeune (2009), Fraistat and Jones (2009), and Siemens et al. (2010). On the other hand, due to a change of paradigm our architectural proposal does not require the complexity of TextGrid as described by Zielinski et al. (2009). More recent research work raises the need to have several views of the encoding (Brüning et al., 2013). In our approach different views are also relevant for interoperability and to simplify the implementation of user interfaces. The work of Wittern (2013) stresses the need to allow dynamic edition of texts and management of versions.

The specific correlation of static and dynamic goals in the *LdoD* Digital Archive means that our emphasis falls on open changes that feedback into the archive. The TEI encoding and software design implications of this project make us address both conceptual aspects of TEI schemas for modelling texts and documents, and the processing problems posed by user-oriented virtualization of Pessoa's writing and bibliographic imagination.

During the conference we intend to make a more detailed presentation of the *LdoD* Archive and show a demo of the prototype being developed.

Acknowledgment

We would like to thank Timothy Thompson for his contributions to the TEI template for *LdoD* and Diego Giménez for the encoding of *LdoD* fragments.

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under projects PTDC/CLE-LLI/118713/2010 and PEst-OE/EEI/LA0021/2013.

Bibliography

- Barney, Brett (2012). 'Digital Editing with the TEI Yesterday, Today, and Tomorrow', in Textual Cultures, 7.1: 29-41.
- Benel, Aurelien and Lejeune, Christophe (2009). 'Humanities 2.0: Documents, Interpretation and Intersubjectivity in the Digital Age'. International Journal on Web Based Communities, 5.4: 562-576. DOI:10.1504/ijwbc.2009.028090
- Brüning, Gerrit, Katrin Henzel, and Dietmar Pravida (2013). 'Multiple Encoding in Genetic Editions: The Case of "Faust"', Journal of the Text Encoding Intiative, 'Selected Papers from the 2011 TEI Conference', Issue 4, March 2013. http://jtei.revues.org/697
- Burnard, Lou and Syd Bauman, eds. (2012). TEI P5: Guidelines for Electronic Text Encoding and Exchange, Charlottesville, Virgina: TEI Consortium. Available at http://www.tei-c.org/Guidelines/P5/
- Earhart, Amy E. (2012). 'The Digital Edition and the Digital Humanities', in Textual Cultures, 7.1: 18-28.
- Fraistat, Neil and Jones, Steven (2009). 'Editing Environments: The Architecture of Electronic Texts'. Literary and Linguistic Computing, 24.1: 9-18. DOI: 10.1093/llc/fqn032
- Schlitz , Stephanie and Bodine, Garrick (2009). 'The TEIViewer: Facilitating the Transition from XML to Web Display'. Literary and Linguistic Computing, 24.3: 339-346. DOI: 339-346.doi: 10.1093/llc/fqp022
- Schmidt, Desmond and Colomb, Robert (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.
- Siemens, Ray, Mike Elkink, Alastair McColl, Karin Armstrong, James Dixon, Angelsea Saby, Brett D. Hirsch and Cara Leitch, with Martin Holmes, Eric Haswell, Chris Gaudet, Paul Girn, Michael Joyce, Rachel Gold, and Gerry Watson, and members of the PKP, Iter, TAPoR, and INKE teams (2010). 'Underpinnings of the Social Edition? A Narrative, 2004-9, for the Renaissance English Knowledgebase (REKn) and Professional Reading Environment

(PReE) Projects', in Online Humanities Scholarship: The Shape of Things to Come, edited by Jerome McGann, Andrew M Stauffer, Dana Wheeles, and Michael Pickard, Houston, TX: Rice University Press. 401-460

- Tummarello , Giovanni, Morbidoni, Christian, and Pierazzo, Elena (2005). 'Toward Textual Encoding Based on RDF'. Proceedings of the 9th ICCC International Conference on Electronic Publishing. http://elpub.scix.net/data/works/att/206elpub2005.content.pdf
- Vanhoutte, Edward (2006). 'Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text-Encoding for Electronic Editions', in Electronic Textual Editing, edited by Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, New York: Modern Language Association of America. 161-180.
- Wittern, Christian (2013). 'Beyond TEI: Returning the Text to the Reader', Journal of the Text Encoding Intiative, 'Selected Papers from the 2011 TEI Conference', Issue 4, March 2013. http://jtei.revues.org/691
- Zielinski, Andrea, Wolfgang Pempe, Peter Gietz, Martin Haase, Stefan Funk, and Christian Simon (2009). 'TEI Documents in the Grid'. Literary and Linguistic Computing, 24.3: 267-279. DOI: 10.1093/llc/fqp016

TEI <msDesc> and the Italian Tradition of Manuscript Cataloguing

Trasselli, Francesca; Barbero, Giliola; Bagnato, Gian Paolo[17]

The Central Institute of Cataloguing (ICCU - Istituto Centrale per il Catalogo Unico e per le informazioni bibliografiche) of the Italian Ministry of Heritage and Culture uses the Text Encoding Initiative standard in the exchange of the manuscripts descriptions processed with Manus OnLine (http://manus.iccu.sbn.it/). Manus OnLine is the Italian national manuscript cataloguing project and at the same time it is the name of a widespread cataloguing software, used by more than 420 people among librarians and researchers. The catalogue contains around 130,000 files that are created using a web application that deals with a relational database in MySQL. The whole software is open-source based.

The current web application allows the sharing of the authority file (which is a rich index of names involved with the manuscripts), and includes some tools that make it more agile in being able to insert and edit the manuscript descriptions. In the software four years of life, between 2009 and 2013, the cooperative work has proved to be very useful and above all procedures were streamlined for the publication of the manuscript descriptions within the OPAC which, in turn, has become an important tool continuously being updated -- a real and proper catalogue in progress. But, in spite of the validity and importance of this cooperative catalogue, some individual libraries and projects that operate simultaneously in different institutions of conservation, need to treat their data outside the central DB. These operators have asked for the export of their manuscript descriptions, because in most cases they want to handle them independently in digital libraries. ICCU has then chosen to create an automatic tool that produces valid TEI documents. This choice respects the need to distribute the

[17] Paper written by Francesca Trasselli, as coordinator of the ICCU's Area of activity for the bibliography, cataloging and inventory of manuscripts, in collaboration with Giliola Barbero and Gian Paolo Bagnato who have respectfully researched and ultimately realized the exportation procedure.

processed data to libraries that produced it, and continue to exercise the right over it.

In December 2012, a new module was added to the software, which allows the export of all the descriptions of a project, a library, a specific collection or even the description of a single manuscript. The new software module was studied by Giliola Barbero and Gian Paolo Bagnato in collaboration with the Area of activity for the bibliography, cataloging and inventory of manuscripts.

The choice of the TEI schema was made after careful consideration of bibliographic standards based primarily on the International Standard for Bibliographic Description (ISBD), that is to say MARC and UNIMARC, given that many colleagues had initially expressed a preference for a common standard shared both by cataloguers of manuscripts and printed publications. However, the assessment of MARC and UNIMARC has led to negative results. Although they are used by some libraries for the structuring of their manuscript descriptions, they do not in fact cover the typical information of the manuscript description and above all the macrostructure of such. Manuscript cataloguing has been traditionally done by first creating a description of the physical aspects of the manuscript, and then the description of a variable number of texts. In case of composite manuscripts, the cataloguing proceeds by creating a description of certain physical aspects shared by the entire manuscript, then the description of the physical aspects of the parts composing the manuscript and, finally, the description of a variable number of texts.

This paper will first demonstrate the relevance of the TEI schema with respect to this traditional macrostructure by showing how it coincides with the most significant in the history of cataloguing of manuscripts. Therefore, the points of contact between the elements used in msDesc of the TEI schema, the UNIMARC and the Dublin Core will be highlighted, and we will attempt to provide a mapping of key information shared by all the three standards.

Secondly, this paper will discuss some critical aspects of the standards and how these have been temporarily resolved. These critical points mainly concern the following elements and information that do not always result in being suited for structuring:

- supportDesc
- extent
- measure
- technical terms in the binding description
- technical terms in the music notation description
- information on manuscript letters

The ICCU evaluated the solutions adopted by the European e-codices and Manuscriptorium projects to describe the support, the number of folios and size of the manuscripts (solutions that differ among each other) and chosen to adapt to the most suited practice in line with the needs of Manus OnLine. However, it has avoided creating further diversification, and currently, it believes that a common choice would be useful. As it regards the binding description and the music notation, while having exploited the element term of the TEI schema, the ICCU believes that it would be necessary to reflect further. It is also absolutely necessary to delve into and discuss the encoding of the physical description and content of manuscript letters in strict accordance with the components of the element msDesc.

Bibliography

- G. Barbero, S. Smaldone, Il linguaggio SGML/XML e la descrizione di manoscritti, «Bollettino AIB», 40/2 (giugno 2000), 159-179.
- Reference Manual for the MASTER Document Type Definition. Discussion Draft, ed. by Lou Burnard for the MASTER Work Group, revised 06.Jan. 2011: http://www.tei-c.org/About/Archive_new/Master/Reference/oldindex.html
- T. Stinson, Codicological Descriptions in the Digital Age, in Kodikologie und Paläeographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age, hrgb. Von /ed. by M. Rehbein, P. Sahle, T. Schaßan, Norderstedt, BoD, 2009, 35-51.
- Zdeněk Uhlíř, Adolf Knoll, Manuscriptorium Digital Library and Enrich Project: Means for Dealing with Digital Codicology and Palaeography, in Kodikologie und Paläeographie, 67-78.
- P5: Guidelines for Electronic Text Encoding and Interchange: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

- e-codices - Virtual Manuscript Library of Switzerland: http://www.e-codices.unifr.ch/en
- Manuscriptorium: http://www.manuscriptorium.com/
- Manus OnLine: http://manus.iccu.sbn.it/

A stand-off critical apparatus for the libretto of Der Freischütz

Viglianti, Raffaele; Schreiter, Solveig; Bohl, Benjamin

Digital editions of opera librettos have been prepared using TEI in several occasions; notable examples are *Opera Liber* [18] (Pierazzo 2005) and OPERA[19] (Münzmay et al. 2011). *Opera Liber* publishes critical editions of librettos with the aim of promoting them as literary text worthy of scholarly attention, in contrast to the common perception of librettos as ancillary material to operatic works. OPERA, on the other hand, develops around the premises that libretto and score are edited according to two independent traditions and moves first steps towards an integrated edition of libretto and music sources.

The BMBF-funded project *Freischütz Digital* (*FreiDi*)[20] takes a broad approach on the matter, with work packages dedicated to the digitization of different kinds of sources of Carl Maria von Weber's opera *Der Freischütz*. The project will include encoded text of both libretto sources (in TEI) and score sources (in MEI)[21], as well as recorded audio

[18] cf. "Opera Liber - Archivio Digitale Online Libretti d'Opera:" available at: http://193.204.255.27/operaliber/.

[19] cf. "OPERA – Spektrum des europäischen Musiktheaters in Einzeleditionen" available at: http://www.opera.adwmainz.de/index.php?id=818.

[20] cf. "Freischütz Digital. Paradigmatische Umsetzung eines genuin digitalen Editionskonzepts." available at: http://freischuetz-digital.de.

[21] cf. "MEI. The Music Encoding Initiative" available at: http://www.music-encoding.org/.

performances. Some of the modelling challenges for this project include minimizing redundancy throughout the encoding, coordinating the corpus and modelling variance and editorial intervention across the material. This paper discusses the approach taken to model the critical apparatus for the libretto, which uses stand-off techniques to encode variance across the corpus and aims at being able to refer to both textual and musical sources.

Sources

There are several sources for *Der Freischütz* libretto and most are easily accessible. They show that the work changed significantly over a long period of time, from first ideas of the librettist Friedrich Kind (1817) to the premiere of Weber's opera on 18 June 1821. Moreover, they reveal that Weber himself was crucially involved in the writing process. Proof for this can be found in the surviving manuscript and printed sources: the manuscript of the librettist Friedrich Kind, Weber's manuscript copy, the surviving copies of the textbook in Berlin, Vienna (KA-tx15), and Hamburg, as well as the first print of the songs, the latter missing the dialog passages. Weber's autograph score (A-pt), several score copies, and the printed piano reduction constitute a corpus of revealing comparative sources to the libretto sources. Moreover, multiple printed editions that Kind published from late 1821 / early 1822 to 1843 – all of which were meant as reading editions – show even more text versions and variants. Weber first sent manuscript copies of the libretto to a few theatres, but later sent the first complete print edition, which significantly influenced the performance tradition and reception of the work.

Common critical editorial practice in music balances historical overview with performance practice and produces "performable" texts, which often are a highly hypothetical construct based on an amalgamation of sources. In this context, the benefit of a digital edition is to transparently depict textual evolution and facilitate the mutually informed investigation and presentation of music and text sources.

Model

FreiDi includes a TEI encoding for each of the libretto sources listed above. The encoding focuses on the dramatic and lyrical structure of the texts, while preserving original spelling, deleted and added material, etc.

These independent transcriptions are coordinated through a collation-like file (a "core" file) that encodes textual variance with `<rdg>` elements containing pointers to markup in the encoding of the sources. In general, this approach is similar to collations generated after an alignment step in modern collation software such as Juxta and CollateX[22]; however, it is designed to operate at more than one level of tokenization, so that statements about variation can be attached to any element in the TEI-encoded sources. Similarly to the 'double-end-point-attached' method, the "core" file allows to address variants that would cause overlapping issues when encoded with the 'parallel segmentation' method;[23] yet it differs from it by keeping `<app>` statements independent from each other and from the text[24]. This approach is motivated by the fact that not every difference between sources will be marked as a variant, such as different uses of the *Eszett* or differences due to document structure such as patches and paste-overs. Using the core file to only identify what are considered "meaningful" variants allows the transcription to keep a higher level of detail without creating issues for collations.

The transcriptions focus substantially on the encoding of the dramatic structure; in fact, the data model will not use the new genetic encoding module since it imposes an important paradigm switch from a text-focused to a document-focused encoding. The editors, nonetheless, can still be detailed about their transcription partly because variation statements are kept separately.

To briefly illustrate this model, let us consider the following verses from source KA-tx15 and A-pt and the corresponding core file entry.

Source KA-tx15.xml	**Source A-pt.xml**

`<l xml:id="KA-tx15_l1">Sie erquicke,`
`</l>`
`<l xml:id="KA-tx15_l2">Und bestricke`
`</l>`
	`<l xml:id="A-pt_l1">Sie erquicke,`
	`</l>`
	`<l xml:id="A-pt_l2">und beglükke`
	`</l>`

[22] See for example the page about "Textual Variance" on the TEI Wiki: http://wiki.tei-c.org/index.php/Textual_Variance#Aligner.

[23] See Chapter 12 of the TEI Guidelines: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPLK.

[24] As such, this model also differs from Schmidt and Colomb 2009, although it shares the approach of not mixing encoded sources and editorial statements to avoid overlapping hierarchies.

```
<l xml:id="KA-tx15_l3">Und beglücke,
</l>
```

```
<l xml:id="A-pt_l3">und bestrikke.
</l>
```

Core

```
<app>
<rdg wit="#KA-tx15">
<ptr target="KA-tx15.xml#KA-tx15_l2"/>
<ptr target="KA-tx15.xml#KA-tx15_l3"/>
</rdg>
<rdg wit="#A-pt">
<ptr target="A-pt.xml#A-pt_l2"/>
<ptr target="A-pt.xml#A-pt_l3"/>
</rdg>
</app>
```

In this example, the core file records the inversion of verses and the <app> statement is limited to a verse-level domain. The core is made of independent <app> statements, so that differences in capitalization, punctuation and spelling that are not included at this point are encoded as separate statements instead. To record this, the granularity of encoding needs to be greater as shown in the following example.

Source KA-tx15.xml

```
<l xml:id="KA-tx15_l1">Sie erquicke,
</l>
<l xml:id="KA-tx15_l2">Und
<w xml:id="KA-tx15_w1">bestricke</w>
</l>
<l xml:id="KA-tx15_l3">Und beglücke,
</l>
```

Source A-pt.xml

```
<l xml:id="A-pt_l1">Sie erquicke,
</l>
<l xml:id="A-pt_l2">und beglükke
</l>
<l xml:id="A-pt_l3">und
<w xml:id="A-pt_w1">bestrikke.
</w></l>
```

Core

```
<app>
<rdg wit="#KA-tx15">
<ptr target="KA-tx15.xml#KA-tx15_w1"/>
</rdg>
<rdg wit="#A-pt">
<ptr target="A-pt.xml#A-pt_w1"/>
</rdg>
</app>
```

Discussion

Pointing to TEI sources from the “core” file introduces the managerial complexity typical of stand-off markup; for example pointers need to be validated and verified. These issues can be overcome by efficient project management and good authoring tools. The model, however, requires that the TEI-encoded sources include semantically weak elements such as <seg>, <w>, <c> and <pc> in the sources, whose only role is to allow

the core file to refer to the text at the right point. Managing this elements is considerably more laborious than managing id references. It would be more efficient to be able to point (or *annotate*)[25] portions of text without needing further XML elements. The TEI XPointer schemes may be useful in this case:[26]
Core

```
<app>
<rdg wit="#KA-tx15">
<ptr target="string-range(xpath1(*[@xml:id='KA-tx15_l2'],4, 9))"/>
</rdg>
<rdg wit="#A-pt">
<ptr target="string-range(xpath1(*[@xml:id='A-pt_l3'], 4,9))"/>
</rdg>
</app>
```

However, implementations of XPointer are currently uneven and limited to XInclude, so using this approach in *FreiDi* would require to implement the schemes. Moreover, the current definition of *string-range()* operates within a "fragment", or a well-formed XML context. This would make it difficult to select ranges that include an opening or closing tag. Hugh Cayless (2012) has recently suggested that TEI XPointer ought to be more sophisticated and proposed an extension of the schemes.
Finally, the model has also been designed to classify `<app>` statements according to a specific taxonomy; this results from keeping the statements separated, so that they address sections of text at different, possibly overlapping, levels. Categorizing variants has been one of the topics of discussion within the Manuscript Special Interest Group, which has been working on a revision of the critical apparatus module.[27]The discussion around categorization has focused on what variants address, such as omissions, punctuation, transpositions. etc. The *FreDi* project team is considering differentiating between variants addressing spelling,

[25] Thinking of apparatus entries as annotations means that other standards specific to annotation may be used in this scenario, for example the Open Annotation Collaboration model: http://www.openannotation.org/.

[26] See Chapter 16 of the TEI Guidelines: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SATS. *xpath1()* applies an XPath expression to select an element, while string-range() identifies a textual range within the selected element, for example starting from position 4 and including the following 9 positions.

[27] A report of a recent workgroup meeting is available on the TEI Wiki: http://wiki.tei-c.org/index.php/Critical_Apparatus_Workgroup.

punctuation, transposition and variance caused by setting the text into music. The latter category in particular has not yet been explored in the field of digital editing.

Conclusions

FreiDi is an ambitious project that handles numerous sources, both musical and literary. This is in line with modern approaches to opera editing, which acknowledge that limiting investigation to only the score or only the libretto is not a desirable approach (Strohm 2005). As a general approach, editorial statements are encoded separately from the sources, with the aim of keeping independence between the source encodings and reduce redundancy. Concerning the libretto, a primarily literary form, a stand-off TEI "core" file is designed to handle the critical apparatus and similar cross-source editorial statements. This organization allows one to organize the statements according to a taxonomy, a feature that has been on the wish-list of the Manuscript SIG for a while. The core file relies on being able to point to specific portions of the TEI source, and techniques that implement this are still being perfected by the community. This project aims at contributing to research in these TEI-related aspects, as well as contributing to the debate around digital editions of operas.

Bibliography

- Cayless, H., 2012. TEI XPointer Requirements and Notes (Draft). Available at https://docs.google.com/document/d/1JsMA-gOGrevyY-crzHGiC7eZ8XdV5H_wFTlUGzrf20w
- Münzmay, A. et al., 2011. *Editing Opera: Challenges of an Integrated Digital Presentation of Music and Text based on "Edirom" and TEI*. TEI Members Meeting 2011, Universität Würzburg, 10-16 October.
- Pierazzo, E., 2005. *An Encoding Model for Librettos: the Opera Liber DTD*. ACH/ALLC 17th Joint International Conference, University of Victoria, British Columbia, 15-18 June.
- Schmidt, D. and Colomb, R, 2009. 'A data structure for representing multi-version texts online'. International Journal of Human-Computer Studies , 67.6, 497-514.

• Strohm, R., 2005. 'Partitur und Libretto. Zur Edition von Operntexten'. *Opernedition. Bericht über das Symposion zum 60. Geburtstag von Sieghart Döhring*, ed. Helga Lühning and Reinhard Wiesend. 37-56.

Panels

Computer-mediated communication in TEI: What lies ahead

Beißwenger, Michael; Lemnitzer, Lothar

Introduction

The social web has brought forth various genres of interpersonal communication (*computer-mediated communication*, henceforth: *cmc*) such as chats, discussion forums, wiki talk pages, Twitter, comment and discussion threads on weblogs and social network sites. These genres display linguistic and structural peculiarities which differ both from speech and from written text. Projects that want to build and exchange cmc corpora would greatly benefit from a standard that allows the user to annotate these peculiarities in TEI.

From the perspective of several corpus projects which aim at building and annotating cmc corpora for several European languages, this panel will discuss how the models provided by the TEI encoding framework may be adapted to the special requirements of cmc genres.

The basis of the discussion is a customized TEI schema presented at the TEI conference held in Würzburg 2011 (Beißwenger et al. 2012)[28]. The panel papers will elaborate on basic features that a TEI standard for cmc resources should include and outline open issues with which further work will have to deal.

The overall goal of the panel is to stimulate the discussion within the TEI community about how a standard for the representation of cmc in TEI should look like and what might be a practical and reasonable way to go about creating such a standard.

In order to push the development of a general standard for the representation of cmc genres and cmc discourse forward, the papers in the panel will present problem overviews for basic issues in representing cmc features in TEI P5 and outline perspectives as well as first suggestions for the treament of these challenges through modifications and expansions of the encoding framework. Starting from these suggestions, the group is

[28] The ODD document can be found at http://www.empirikom.net/bin/view/Themen/CmcTEI

planning to work out feature requests and load them onto the TEI projects page on sourceforge.net.
After a general introduction, paper 1 asserts that solutions for the representation of cmc in TEI should be included in the official TEI guidelines and not remain a task that research and corpus projects have to solve using individual customizations. In addition, the paper formulates general requirements a framework for the representation of cmc (in TEI) should comply with as well as specific requirements from several projects which are currently building corpora of cmc discourse for four European languages (Dutch, German, French, and Italian).
Taking into account the requirements outlined in paper 1, paper 2 starts wich an overview of existing suggestions for the representation of basic structural and linguistic features of cmc discourse in the TEI framework. It then presents considerations on the following open issues: (1) the modeling of different types of citations in cmc postings; (2) the modeling of hypermedia features (hyperlinks and linking structures, embedded media objects); (3) challenges related to the representation of discourse in multimodal cmc environments in which the participants in one interaction space combine a variety of modalities from written, spoken and non-verbal modes.
Paper 3 examines the issue of metadata. It discusses general requirements for representing metadata of cmc resources and outlines a proposal for representing cmc metadata in the TEI framework.
The panel will include 30 minutes of discussion time (15 minutes each after paper 2 and 3).

Paper 1: Modeling computer-mediated communication in TEI: requirements and perspectives

Michael Beißwenger; Thierry Chanier; Isabella Chiari; Maria Ermakova; Lothar Lemnitzer; Angelika Storrer; Maarten van Gompel; Henk van den Heuvel

This paper reports an ongoing work in a network of corpus projects which aim at building and annotating corpora of computer-mediated communication (cmc)[29] and asserts that a framework for the

[29] http://wiki.itmc.tu-dortmund.de/cmc/

representation of cmc should become a part of the TEI guidelines. It gives an overview of research fields in the Humanities and Computer Sciences which would benefit from the availability of such a representation framework and outlines the basic requirements it will have to comply with:

- The schema should provide a general model for the description of the structural and linguistic peculiarities of cmc discourse.
- To be useful for a broad range of application contexts in the Humanities, it should not be designed with one single project in mind but it should take into account the specific requirements of several projects (and genre typologies) in which the creation of annotated cmc resources is of interest.
- In order to be suitable for small data sets which are annotated manually and also for the annotation of big data (e.g., reference corpora in Linguistics, large web corpora in the field of Natural Language Processing), its basic structure should be defined in a way that favours or supports (at least partially) automatic annotation procedures.
- The schema should build on a review of models which already exist in the TEI framework (currently TEI P5) and adapt them to the peculiarities of cmc genres in a reasonable and practical way.
- It should reflect the fact that CMC shares characteristics with written text as well as with spoken conversation while at the same time it is significantly different from both in its textual form and in the mode of production and reception.
- It should allow for an easy (and reversible) anonymization of cmc resources for purposes in which they shall be made available for other researchers (e.g., in the case of reference corpora).
- It should allow for an easy referencing of random samples of the resource (e.g., for citation in scientific publications, didactic materials or dictionary articles).

Since papers 2 and 3 of the panel take into consideration the goals and needs of several projects which are currently dealing with the construction of corpora of cmc discourse in four European languages, paper 1 includes a brief presentation of the four projects and an outline of their project-specific requirements for an annotation schema:

- *DeRiK* ("Deutsches Referenzkorpus zur internetbasierten Kommunikation") is a joint project of TU Dortmund University and the Berlin-Brandenburg Academy of Sciences (BBAW) and the Humanities which is building a reference corpus of German cmc discourse including the most prominent cmc genres. The DeRiK corpus will form a new component of the reference corpora of contemporary written German collected in the BBAW project "Digitales Wörterbuch der deutschen Sprache" (DWDS). On the one hand, it is designed as a resource for corpus-based linguistic analyses of language use in German cmc as well as – in combination with the DWDS corpus – of the impact of cmc genres on contemporary written German. On the other hand it will serve as a resource for the lexicographic description of "netspeak" vocabulary and cmc-specific processes of lexical-semantic change in the dictionary component of the DWDS online lexical information system[30] (cf. Beißwenger et al. 2013). For annotation, DeRiK is currently using the customized TEI schema for cmc described in Beißwenger et al. (2012). The schema comprises, among others, an element for the description of user contributions to cmc conversations (the divLike element *posting*), a distinction of two major types of cmc macrostructures (the cmc-specific division types 'thread' and 'logfile'), a component for modeling the authors of cmc postings as well as elements for the annotation of selected "netspeak" features in individual user postings (emoticons, interaction words, interaction templates, addressing terms).
- The Dutch reference corpus *SoNaR* was intended to serve as a general reference for studies involving language and language use. The corpus should provide a balanced account of the standard language and the variation that occurs within it. In doing so, it allows researchers investigating language use in a particular domain (e.g. medicine) or register (e.g. academic writing) or by a specific group (e.g. professional translators) to relate their data and findings to the general reference corpus. The corpus

[30] http://www.dwds.de

was also intended to play a role in the benchmarking of tools and annotations. Collected in 2008-2012 the corpus contains 500 Mwords, including discussion lists, e-magazines, websites, Wikipedia, SMS, chats and tweets. SoNaR is delivered in the FoLiA format (van Gompel 2012). FoLiA aims to support a wide variety of linguistic annotations in a generic paradigm and has been successfully adopted by various projects in The Netherlands. To provide support for new media, a type of structure annotation called "event annotation" was added, which fits nicely in the paradigm. SoNaR incorporates support for tweets, chat logs and SMS. The former two have been encoded as events, in which each tweet or chat message constitutes an event. Within the event structure, further subdivisions can optionally be made, such as paragraphs, sentences, words (in case of tokenized data). Elements in FoLiA carry a class from a certain set. In this way flexibility is provided to the user. The sets can be formally defined. The events in SoNaR are assigned classes such as "tweet" or "chatmessage". The actors of the set are also explicitly annotated, and further metadata on the annotation is also supported.

- *LETEC* ("Learning & Teaching Corpora"). Mulce repository[31] is a databank of LETEC corpora built upon online learning situations (Reffay, Betbeder & Chanier, 2012). All interactions among participants have been collected and structured before their analysis. It assembles a large variety of cmc types: email, forums, chat, blogs, 3D environments with audio and text chats, etc. One of the main components of its XML structure (Mulce-struct)[32] is the *workspace*. It includes descriptions of its *members* as *references* to the participants registered in the learning activity, *starting* and *ending dates*, the *tools* and the interaction tracks or acts that occurred using these tools. Each cmc tool has a detailed and specific structure. Large subparts of the LETEC databank will be integrated in 2013-14 into a nationwide cmc corpus in French where other

[31] http://repository.mulce.org

[32] Schema for the instantiation component of a LETEC corpus. http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/mce_sid.xsd6

cmc types, such as SMS, tweets, Wikipedia forums, will be added. The cmc SIG group leading the project belongs to the national consortium *"IR corpus-écrits"* in charge of building a reference corpus in French. The cmc SIG has designed a working package which will take care of the cmc TEI structure[33] of the whole corpus and work jointly with the European colleagues gathered in this panel.

- *Web2Corpus_it* ("Corpus italiano di comunicazione mediata dal computer") is a project funded by Sapienza University of Rome in 2010 aimed at investigating meaning negotiation strategies in cmc. It focuses on conversational, interactive, public, written communication in order to build a genre-balanced cmc corpus of Italian language to be investigated both qualitatively and quantitatively. The genres included are: forum, blog, newsgroup, social network and chat (cf. Chiari and Canzonetti, in press)[34]. The collected corpus comprises one million words and has been fully anonymized (by masking), in order to avoid personal details of participants being disclosed, and xml-annotated both for macro-structural properties (thread, post, sender details - avatar | signature | nickname | senderplace – subject, date, time, links and embedded media, web action elements and cmc-specific emoticons and tags and addressing terms). At present the corpus is being processed linguistically with a statistical POS tagger and lemmatizer, including a reference machine dictionary (Common Lexicon of Italian) developed in order to include cmc specific lexical items, and will be subsequently manually checked and is planned to be released in late 2013.

These four corpus projects will provide the test bed for an evaluation of the models under construction with cmc discourse from different languages.

[33] https://groupes.renater.fr/wiki/corpus-ecrits-nouvcom/public/proj-tei/index

[34] http://www.glottoweb.org/web2corpus/

Paper 2: Expanding the TEI encoding framework to genres of computer-mediated communication: considerations and suggestions

Michael Beißwenger; Thierry Chanier; Isabella Chiari; Maria Ermakova; Lothar Lemnitzer; Angelika Storrer; Maarten van Gompel; Henk van den Heuvel

The first section of this paper presents some basic suggestions for the expansion of the TEI encoding framework to the structural and linguistic particularities of cmc genres. It takes into account the general requirements as well as the project-specific requirements outlined in paper 1 and builds on the customized TEI schema for cmc which has been presented at the 2011 TEI members' meeting (published in Beißwenger et al. 2012). The suggestions describe features for the modeling of corpus documents with stored discourse from cmc genres such as online forums, chats, wiki talk pages, Twitter, weblogs or social network sites and (amongst others) refer to the following basic issues in the description of cmc:

- the representation of user postings in written cmc as units which share characteristics with both text and conversations: under aspects of planning and coherence, they are designed as moves in an ongoing conversation; under the aspect of production and reception they behave just like texts, which first have to be produced and then are presented to and received by the addressee(s) *en bloc*;
- the need for models for the representation of *cmc macrostructures* (= the way how series of user postings are grouped / presented to the users, e.g., in the form of *logfiles*, different types of *threads*, *timelines* etc.);
- the need for elements for the annotation of cmc-specific structural and linguistic features on the *microlevel of cmc discourse* (= the content of the postings which comprises e.g. typical "netspeak" phenomena such as emoticons, action words, addressing terms; hashtags; speedwriting phenomena, phenomena

of non-standardized writing; embedded hyperlinks and media objects etc.);

With the help of examples from the corpus projects introduced in paper 1, the second section of the paper will offer problem sketches of the following open issues in modeling cmc and outline some first ideas for their treatment in TEI:

- *Handling citations:* Especially in forums and Bulletin Boards, cmc postings often contain (simple and nested) citations which reproduce content that has originally been part of other authors' prior postings. A schema for the representation of cmc should include a model for the annotation of citations and for referencing citations with the cited prior postings and their authors.
- *Cmc data as hyperlinked data:* Many cmc resources contain hyperlinks and linking structures. A framework for the representation of cmc interactions must include models for the description of how postings are linked with each other and/or with other interaction-external resources on the internet. In some cmc applications (e.g., micro-blogging sites such as Twitter) the method of displaying one and the same user posting as part of a sequence may vary depending on the user's choice (cf. e.g. on Twitter the timeline of one author's tweets vs. the timeline of tweets by different authors which include the same hashtag). A general model for cmc resources must provide features for the description of these kinds of structures and of the target sources of the hyperlinks.
- *Dealing with data from multimodal cmc environments:* In some cmc environments users are communicating not only in a text-based mode but using a combination of text-, audio-, video- and/or 3D-based modalities of interaction (e.g., e-learning platforms, *Skype*, gaming environments, virtual worlds etc.). One of the challenges related to the representation of cmc discourse recorded in environments of that kind is that contributions created and sent in one modality may contribute to, and indeed supplement, a contribution in another modality. In audio-graphic conferencing environments such as Skype, written postings sent via chat may contribute to an ongoing spoken conversation in the audio modality.

In collaborative writing environments, written postings in the chat may contribute to the creation of a longer stretch of text in the word-processing modality. One challenge of treating cmc discourse of that kind is thus the necessity to integrate and align user contributions made in different modalities into a representation of the overall *multimodal* interaction. Since TEI provides modules not only for written but also for (transcriptions of) spoken discourse, the different modes could be represented separately (using different TEI modules) while the alignment of the utterances and postings in the different modalities would have to be solved in an additional representation which is connected with the different resources.

Paper 3: Metadata for cmc documents

Axel Herold; Lothar Lemnitzer; Michael Beißwenger; Isabella Chiari

Extensive and correct metadata has been recognized to be a crucial property of every data object that is used as a primary data source in research contexts. Fine grained metadata allow for *identification, location* and *management* of resources (e.g., NISO, 2004) but also provide researchers with crucial information regarding the *suitability* of a given resource for their particular research interest. The TEI header recognizes all of these metadata requirements to different degrees (Burnard 2005).
Our paper will have a strong focus on the encoding of intrinsic properties of different cmc data sets, thus addressing the issue of finding resources which are suitable for a given research question. Ideally, this part of the metadata description is based on the model representing the primary data. In this respect our paper strongly relies on paper 2, which will propose such a model for cmc data.
An example of cmc-specific data types are emoticons: small, iconic representations of an interlocutor's emotion or his/her attitude towards an utterance (either self produced or produced by other speakers) or towards a communication peer, to name just some of their communicative functions. It is therefore worth considering to either encode normalization and classification schemes for those entities within the metadata description or to provide pointers to such schemes in addition to a suitable markup of these entities within the primary data.

Cmc data often contains large portions verbosely cited material from previous parts of the discourse. This creates a challenge to the measurement of the extent of a given resource. Depending on the assumed discourse status of cited (parts) of utterances it may be necessary to include or exclude cited material. This is a theory-dependent decision, and it should therefore be possible to give concurrent values for a single unit of measurement. Moreover, metadata information on (the handling of) citations may – to some extent – be derived from the primary data directly (see paper 2 for handling of citations in primary text).

Distinct ypologies for cmc tools (including tools that were used to access the primary data) and cmc genres are needed to account for the broad range of different data sources, e.g., online forums, chats, wikis, Twitter, weblogs, social network sites, learning environments and others. We will suggest mechanisms of referencing a particular typology of cmc genres from within the metadata, however, without making any regulations on which kind of typology should be used and referenced in a given project.

Special care must be taken in the metadata description of information about discourse participants to ensure privacy and/or anonymity of the speakers involved in the discourse. Moreover, specific metadata for cmc should also have the function of restoring context information about features of the communication mode of production and reception of cmc texts that are not evident in the text itself. This involves features such as the temporal structuring of the discourse (synchronous vs. asynchronous mode), conversational hierarchies among discourse participants (e. g. blog author vs. commentator), discourse topic/domain or accessibility of the discourse (e. g. private vs. closed vs. public). The availability of social and other context information varies greatly, not only in quantity but also in its quality, according to the primary data source. Therefore a cmc metadata scheme will have to account for different levels of reliability for such information.

Considering the given fourfold structure of the TEI header (file description, encoding description, text profile and revision description), we will identify and discuss different possibilities for recording metadata properties that are specific for cmc data:

- Cmc data comprise properties found in traditional written resources (such as books or newspapers) as well as properties found in resources of (transcribed) spoken language. Both types of resources have previously been provided with TEI-based metadata. Properties shared across different resource types can be expected to be reusable for cmc metadata, e.g., listPerson to denote discourse participants or profileDesc to describe general discourse settings.
- Some metadata properties that cannot be readily encoded using specific elements can still be recorded using the generic *feature structure representation* (fs). Embedding of feature structures is currently allowed for a limited set of header elements in the TEI such as *classCode, extent, language, scriptNote* and *typeNote*. Exploiting the semantic linking mechanism provided by *att.datcat* (via the ISOcat data category registry; note that *classCode* provides a native semantic interface via @scheme as well) would allow tailor-made semantics for the properties encoded in such a way. But obviously this adds a level of indirection and does not capture these properties within the TEI directly.
- A third possibility lies in the adaptation of the TEI element inventory or of suggested cmc-specific value sets for existing elements. For individual projects this can already be achieved by TEI customizations but it may hinder interoperability across resources using elements not found in the TEI guidelines – which is another argument for why models for the representation of cmc data in TEI should better be part of the official guidelines and not be something that each project needs to solve individually.

We will conclude the paper with a proposed metadata header for TEI documents encoding cmc data. We will also – at least for some prominent features of metadata for cmc documents – show how the TEI header metadata are related to, and can be converted to, metadata components within the emerging CLARIN Metadata Framework (Component Metadata Infrastructure, CMDI).

Bibliography

- Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative, Issue 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).
- Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2013): DeRiK: A German Reference Corpus of Computer-Mediated Communication. In: Literary and Linguistic Computing (LLC).
- Burnard, Lou (2005): Metadata for corpus work. In: Martin Wynne (ed.): Developing Linguistic Corpora: A Guide to Good Practice. Oxford, 30-46.
- Chiari, Isabella; Canzonetti, Alessio (in press): Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In: Enrico Garavelli & Elina Suomela-Härmä (eds.): Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua. Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI, Helsinki 18-19 June 2012), Franco Cesati Editore, Firenze.
- [NISO 2004] National Information Standards Organization (2004): Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- Oostdijk, Nelleke; Reynaert, Martin; Hoste, Véronique; Schuuman, Ineke (2013): The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In: Peter Spyns & Jan Odijk (eds): Essential Speech and Language Technology for Dutch. Springer. http://link.springer.com/chapter/10.1007/978-3-642-30910-6_13
- Reffay, Christophe; Betbeder, Marie-Laure; Chanier, Thierry (2012): Multimodal Learning and Teaching Corpora Exchange: Lessons learned in 5 years by the Mulce project. Special Issue on dataTEL: Datasets and Data Supported Learning

in Technology-Enhanced Learning. International Journal of Technology Enhanced Learning (IJTEL) 4 (1/2), 11-30. http://edutice.archives-ouvertes.fr/edutice-00718392 (DOI: 10.1504/IJTEL.2012.048310).

- [TEI P5] TEI Consortium (eds) (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5/ (accessed 22 March 2013).
- van Gompel, Maarten (2012). FoLiA: Format for Linguistic Annotation. Documentation. ILK Technical Report 12-03. Available from http://ilk.uvt.nl/downloads/pub/papers/ilk.1203.pdf

The role of the TEI in the establishment of a European shared methodology for the production of scholarly digital editions

Driscoll, Matthew James; Pierazzo, Elena; Buzzoni, Marina; Damon, Cynthia; Burghart, Marjorie; Sahle, Patrick

While it cannot be denied that the TEI represents an important point of reference for the preparation of digital editions of culturally important texts of all kinds, its influence remains somewhat more marginal than should ideally be the case. From an encoding point of view, despite many improvements made in the last few years (see for instance the new mechanisms for documentary and genetic encoding), there are still a few 'grey areas', one of the more obvious being the critical apparatus module (Chapter 12 of the Guidelines), which has several clear gaps and flaws and has been widely criticised in recent years (see e.g. Burghart and Rosselli Del Turco 2012).

More worrying and probably more impactful, is however the lack of easytouse tools supporting the encoding process and the subsequent

management of the encoded files. The question in this case is if these tools are yet to come or whether they will be ever coming (see Pierazzo 2011).

Another major drawback in the general adoption of the TEI by the scholarly editorial community is perhaps represented by the final delivery of the edition once the encoding process is finished. There are a few tools readily available, such as the TEI stylesheets and the TEI Boilerplate, but they are limited, not very easy to customise without specific knowledge and not really suitable for high spec, complex digital editions.

And yet the TEI has undeniably played a vital role in shaping the intellectual agenda with respect to scholarly digital editions. Why does it still meet with resistance from scholars engaged in the production of editions? In 2011 a Europeanwide network called NeDiMAH (Network for Digital Methods in the Arts and Humanities) was launched with the purpose of "carrying out a series of activities and networking events that will allow the examination of the practice of, and evidence for, digital research in the arts and humanities across Europe" (see www.nedimah.eu/). The Network is supported by the European Science Foundation and involves representatives from Bulgaria, Croatia, Denmark, Finland, France, Germany, Ireland, the Netherlands, Norway, Portugal, Romania, Sweden, Switzerland and the United Kingdom. Within NeDiMAH a working group has been set up specifically devoted to Scholarly Digital Editions in seeking to promote international cooperation and to highlight best practices and areas of improvement both in terms of methodologies and IT infrastructure (see http://www.nedimah.eu/workgroups/scholarlydigitaleditions). Following a very successful expert seminar in The Hague (see http://www.nedimah.eu/events/nedimahexpertmeetingdigitalscholarlyeditions) where theoretical and practical issues connected with the production and consumption of scholarly digital editions have been debated, the working group proposes a roundtable specifically focused on the role of the TEI within the theory and practice of scholarly digital editing. The main topics that will be covered are:

The apparatus criticus: How and why? The TEI offers three different formats for encoding variants, but it seems that only "parallel segmentation" has been used in practice by the TEI users. This method has

several drawbacks (for instance, with many witnesses the markup between excessively complex, with much overlapping of lemmas inevitable), but it seems to be the only one that allows for any sort of implementation. The other two methods, on the other hand, in spite of being far more flexible, require a considerable effort in the development of any processing tools, based as they are on standoff markup.
What is really the function of the critical apparatus? The TEI Guidelines seem to imply that it works like a repository of variants. A proper apparatus criticus is far more than that, however: it is the key to understanding why the text presented is what it is. More precisely, the apparatus is a set of notes designed to foster in the reader an awareness of historical and editorial processes that resulted in the text s/he is reading and to give the reader what s/he needs to evaluate the editor's decisions. Is this vision present or even possible within the Guidelines?
Tools: In this context, we will discuss the potential impact of outreach targeting tool developers from outside the strict TEI community. Could we offer developers more or less unfamiliar with the TEI a lowthreshold introduction, less overwhelming than the Guidelines? This would of course require some recommendations for "best practice". Burghart proposes a series of "cheatsheets" (Burghart 2011), offering digests of TEI encoding recommendations starting from the user experience. These could serve not only as a guide to the guidelines for endusers, but could also be of great help to developers to understand the concepts / phenomena their users want to encode.
More generally we will discuss the TEI intellectual leadership and responsibilities in the field of digital scholarly editing.

Participants

- M. J. Driscoll, Københavns Universitet, Chair of the NeDiMAH working group on digital scholarly editions
- Elena Pierazzo, King’s College London, Cochair of the NeDiMAH working group on digital scholarly editions
- Marina Buzzoni, Università Ca' Foscari Venezia
- Marjorie Burghart, L'École des hautes études en sciences sociales, Lyon

- Cynthia Damon, University of Pennsylvania
- Patrick Sahle, Universität zu Köln

Bibliography

- Pierazzo, Elena (2011). 'The Role of Technology in Digital Scholarly Editing'. Paper presented at the TEI Conference and Members' Meeting, University of Würzburg, 1016 October 2011, available from http://www.teic.org/Vault/MembersMeetings/2011/tei_mm_2011/abstracts/abstracts_papers/ind ex.htmlhttp://www.teic.org/Vault/MembersMeetings/2011/tei_mm_2011/abstracts/abstracts_papers/ind ex.html
- Burghart, Marjorie (2011). 'TEI: critical apparatus cheatsheet'. Available from http://marjorie.burghart.online.fr/?q=en/content/teicriticalapparatuscheatsheethttp://marjorie.burghart.online.fr/?q=en/content/teicriticalapparatuscheatsheet
- Burghart, Marjorie and Rosselli Del Turco, Roberto (2012). 'Getting critical with the apparatus: how to rethink the TEI encoding of critical editions?'. Paper presented at the TEI Conference and Members' Meeting, Texas A&M University, 710 November 2012 Avaialble from http://idhmc.tamu.edu/teiconference/program/papers/#editionshttp://idhmc.tamu.edu/teiconference/program/papers/#editions

TAPAS and the TEI: An Update and Open Discussion

Flanders, Julia; Bauman, Syd; Pierazzo, Elena

The TEI Archiving, Publishing, and Access Service (TAPAS) is now entering its second year of development, with the goal of supporting the publication and archiving of small-scale scholarly TEI projects. A

prototype is now being tested which supports a set of core functions including the creation of projects and collections, upload of TEI data, creation of metadata and transfer of metadata from existing TEI files, configuration of the publication interface, and various ways of exploring TAPAS collections. An intensive user testing period is scheduled for the end of April 2013, and an additional period of user testing will be conducted during July and August 2013. An initial release of the service is planned for early 2014. TAPAS is also exploring a relationship with the TEI Consortium that would make TAPAS a benefit of TEI membership, and that would take advantage of TAPAS to offer discounted TEI workshops and supporting services to TEI members.

At the TEI annual conference in 2012 at Texas A&M University, Julia Flanders gave a presentation on TAPAS that sought to elicit ideas and comments from the TEI community concerning the role TAPAS might play in supporting the creation, publication, and long-term archiving of TEI data. The resulting discussion offered input a number of important issues that have had significant impact on the shape of TAPAS: for instance, the suggestion that TAPAS might serve as a kind of community corpus or teaching corpus for TEI data, the issue of divergent encoding practices within TAPAS data, and the question of how to handle migration to future versions of the TEI Guidelines. Following a year of further development, it is important for TAPAS to receive further input from the TEI community and to provide updated information on the project's development.

This session will begin with three short presentations from panelists that offer an updated view of progress on TAPAS, as follows:

1 Julia Flanders will present an update on the technical and strategic development of TAPAS, including the architecture of the service, the business model, and the process of user testing.

2 Syd Bauman will present a detailed examination of the TAPAS schemas and their design, and will report on information gathered through the profiling of TEI data contributed to TAPAS.

3 Elena Pierazzo will present an update on the relationship between TAPAS and the TEI, focusing on the development of

a memorandum of understanding and the planning of TAPAS services as TEI member benefits.

Following these presentations, the session will provide approximately 45 minutes for open discussion. The following questions will be suggested as starting points but any topics raised by audience members will be welcome:

- Can TAPAS be made sustainable as a benefit of TEI membership?
- How can TAPAS better serve the international TEI community? is its scope too limited?
- What are the highest priority features for TAPAS to offer its contributors?
- What are the highest priority features from the reader's perspective? What will make TAPAS a useful resource about the TEI?

Dialogue and linking between TEI and other semantic models

Tomasi, Francesca; Ciotti, Fabio; Lana, Maurizio; Vitali, Fabio; Peroni, Silvio; Magro, Diego

The deep dialogue TEI started with other semantic models – i.e. CIDOC-CRM and FRBR/FRBR (OO) has two aims: the data and documents interchange and the improvement of the editors possibilities to formally declare hermeneutical positions. The TEI schema provides most of the elements/attributes (and classes) useful to describe interpretation instances, while further schemas, as well as other value vocabularies and metadata element sets, are supposed to enhance some potentialities of the model itself. On one hand, additional schemas could contribute to perfect the scope of some TEI elements, while on the other, the existing

ontologies could improve the interpretation effectiveness. Therefore, this panel is aimed at introducing three different approaches to document representation, where TEI may draw some hints from other models.
We first present the contribute of EAC (Encoded Archival Context) to extend people's description, starting from the archival approach to the context, here intended as the key element to define individual's roles and functions. Then we considered the dialogue between TEI and the existing ontologies, with particular attention to geographic data. Finally, thanks to the 'semantic lenses' employed as an exploratory tool for annotated documents, we started up the relationship between TEI and specific ontologies related to semantic publishing.
The aforementioned approaches adopt a linked data perspective, adding the TEI element with @ref and URI and adopting the RDF model for assertions. By exposing TEI annotations as data sets, we could improve both the schema and the documents interchange with other exiting data sets, enhancing the information retrieval possibilities. Digital editions based on TEI could start a dialogue with the WWW resources in a global vision of heritage, here intended as cultural data connection, where digital editions, acting like a sort of interlink between literary texts, archival documents and books, play a crucial role in the preservation of cultural memory.

TEI <person> versus EAC: the identity between functions and context

Tomasi, Francesca

Amongst the most significant changes in the TEI P5 Schema version, the Biographical and Prosopographical Data [1] section undoubtedly constitutes a challenging innovation. TEI decided to invest on 'persons', defining an elements taxonomy useful to describe individuals. In 2006 a special workgroup called 'Personography' was chartered: its task was "to investigate how other existing XML schemes and TEI customization handle data about people" [2] and a "Report on XML mark-up of biographical and prosopographical data" was published [3].
A basic approach to describing people consists in the unique individuals' identification and the description enrichment through features

classification. However, we must never forget that people are strongly connected with the textual context: as a result, roles and functions, intended as individuals' features, naturally change depending on the context, i.e. on the source attesting the individual. It's therefore possible to state that: 1) some features not only are static over the time but they are also theoretically constant in relation to the context (i.e. birth, death, nationality, persName); 2) other features vary depending on date and place (i.e age, affiliation, education, event, state); 3) roles and functions (i.e. author, actor, editor, speaker) are elements that identify people depending on the context.

Thus we can say that a person is a complex entity, because she/he is connected with different phenomena typologies: some are unchangeable, while some depend on a time period, a place or a context. In any case, all these features are able to turn a string into a concept, that is an assertion resulting from the relation between the elements needed to provide meaning.

The <person> element in TEI could be associated with different roles or functions. Let's consider the digital edition of a literary text. We may say that a person is, respectively: the one who created the digital edition - at different levels -, the analogic source author, the printed version editor, the whole of individuals quoted in the text. The concept of person extends its domain: although individuals are strictly related to the source constituting their appropriate semantic background, they are also entities with a function enabling a single person to connect either with different documents - or other resources in general - and several persons with other people sharing the same role. Multiple relationships therefore arise: between individuals, between a person and a document in which she/he is mentioned and between a person and other resources.

This reflection links TEI to one peculiar XML schema, called EAC (Encoded Archival Context) [4] developed in order to formalize the ISAAR (CPF) standard (International Standard Archival Authority Record for Corporate Bodies, Persons and Families)[5] and today represented also as ontology [6]. EAC contributes to the reasoning on individuals, pointing out the importance both of the context and the relationships. The approach here described aims to extend the domain of

digital editions to the archival studies one. The archival science declares the principle of separation between the description of records (documents) and the description of people (corporate bodies, persons and families) [7], focusing on the context as a key element. The same approach could be mostly implemented in TEI, if the final purpose is to expose data sets to be used by the Web community.

It becomes then essential to consider EAC as a schema able to suggest how to extend the concept of <relation> in TEI. EAC (CPF) is based on the principle of entity intended as corporate body, person, or family that manage relationships – between entities and between one entity and a resource linked at some level - each of which could be described, dated and categorized. Besides the elements connected to the "relation" principle (<cpfRelation> and <resourceRelation>), EAC describes the <function> element that "provides information about a function, activity, role, or purpose performed or manifested by the entity being described" [4] on a specific date. The element <functionRelation> describes a "function related to the described entity. [...] Includes an attribute @functionRelationType" that could support a values taxonomy 4].

A new model of authority record, intended as complex structure able to document the context in which the identity is attested, could be introduced: the authority is generated not only by the controlled form of the name, and the related parallel forms, but it is also the result of relationships resulting from the context to determine a concept [8].

According to the RDF model, it's possible to say that an identified entity (URI) manages relationships (predicate) with different objects: another entity (URI), i.e. another person, a place (URI), a date (URI), an event (URI), a contextual resource (URI) i.e. the document, an external resource (URI), that is another object (a document, an image, a video, an audio record, an so on).

We could try to apply this procedure to the responsibility of an individual identifiable as contributor of a digital edition who, on a specific date, performed a specific activity. TEI metadata propose two options for the responsibility description (<fileDesc> e <revisionDesc>): <fileDesc><titleStmt><respStmt> <resp>,<name>

<revisionDesc><respStmt> <resp>,<persName>

Each person is associated with a "responsibility" able to identify the function the entity covered in that document, linking people to resource. The same person could cover the same responsibility in other editions; in this way relationships might be extended to other documents. Other individuals could be moreover connected to the aforementioned person due to the sharing of the same responsibility.

This process could be declared and exposed as data set with RDF and URI for the syntax and TEI/EAC for classes and predicates in order to build a collection of authorities of people who covered either a role or a function in a certain time period and context. By declaring connections as relationships, through the EAC model, we could develop a knowledge base of people, with a context-originated function.

We can definitely say that digital editions open the door to the cultural heritage domain, establishing connections between heterogeneous objects and "creating efficiencies in the re-use of metadata across repositories, and through open linked data resources" [9]. Linked Data describing persons performing specific roles would be considerably improved by employing specifications relative to these persons' function while using the context as interpretative key: "the description of personal roles and of the statuses of documents needs to vary in time and according to changing contexts [...] such roles and statuses need to be handled formally by ontological models." [10]

Bibliography

- [1] TEI Consortium (eds.). "13.3 Biographical and Prosopographical Data". In Guidelines for Electronic Text Encoding and Interchange. Last updated on 21 December 2011. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html#NDPERS
- [2] TEI: Personography Task Force. http://www.tei-c.org/Activities/Workgroups/PERS/index.xml
- [3] Wedervang-Jensen, Eva, and Matthew Driscoll, Report on XML mark-up of biographical and prosopographical data. 16 Feb 2006. http://www.tei-c.org/Activities/Workgroups/PERS/persw02.xml

- [4] EAC-CPF, Encoded Archival Context for Corporate Bodies, Persons, and Families. http://eac.staatsbibliothek-berlin.de/
- [5] CBPS - Sub-Committee on Descriptive Standards. “ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families”. 2nd Edition, 2003. http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html
- [6] Mazzini, Silvia, and Francesca Ricci. 2011. “EAC-CPF Ontology and Linked Archival Data”. In Semantic Digital Archives (SDA) Proceedings of the 1st International Workshop on Semantic Digital Archives. http://ceur-ws.org/Vol-801/
- [7] Pitti, Daniel. 2004. “Creator Description: Encoded Archival Context”. Authority control in organizing and accessing information: definition and international experience. Ed. Arlene G. Taylor, 1941-, Barbara B. Tillett, Murtha Baca and Mauro Guerrini, 201-226. Binghamton N.Y.: Haworth Information Press
- [8] Tomasi, Francesca. 2013. Le edizioni digitali come nuovo modello per dati d'autorità concettuali. JLis 4.2. 10.4403/jlis.it-8808
- [9] Larson, Ray R., and Krishna Janakiraman. 2011. “Connecting Archival Collections: The Social Networks and Archival Context Project”. In Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL 2011). Ed. Stefan Gradmann, Francesca Borri, Carlo Meghini and Heiko Schuldt, 3-14. Heidelberg, Germany: Springer. DOI: 10.1007/978-3-642-24469-8_3
- [10] Peroni, Silvio, David Shotton, and Fabio Vitali. 2012. “Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents”. In Proceedings of the 8th International Conference on Semantic Systems, 9-16. ACM, New York. DOI: 10.1145/2362499.2362502

Geolat: a digital geography for Latin literature

Lana, Maurizio; Ciotti, Fabio; Magro, Diego

This paper presents the "Geolat" project, which aims to make accessible the Latin literature through a query interface of geographic / cartographic type. The project, under the name DAGOCLaT (Digital Atlas with Geographical Ontology for Classical Latin Texts) in 2012 was presented in response to the call of "Compagnia di San Paolo Foundation" and at the end of a blind peer evaluation managed by European Science Foundation was funded for exploratory and initial activities. In January 2013, under the name ALTUSS (Advanced Latin Texts Uses for School and Society) the project, revised and enriched among other things by an advisory board composed by Gregory Crane (Perseus, Pelagios), Tom Elliott (Pleiades) and Leif Isaksen (Google Ancient Places), was presented in response to the European call ERC Synergy.

The first objective of the project is to set up a digital library that contains the works of Latin literature from its origins to the end of the Roman Empire (conventional date, the 476 d. C.). This stage involves the integration of various already existing repository of Latin texts of high philological quality, which will be integrated starting from their already existing TEI/XML encoding. Building a (someone could say "the") global digital library of ancient Latin literature is a very important field where APA is working [1], where Gregory Crane recently called [2] to start working, and where the "Geolat" project too will build its global library, because the library is a pre- condition for all the subsequent activities. All the library texts will be encoded with a very light TEI subset of tags.

In a second phase the works so collected are analyzed at morphological level by means of a parser (that of Lasla of Liège [3]) so as to associate with each word its analysis / morphological description, which includes the identification of proper names. After that, by means of manual intervention, geographic references will be progressively encoded in a formal manner by adopting the TEI elements <placeName> and <geogName> (described in the TEI Guidelines in chapter 13 "Names, Dates, People, and Places"). Each occurrence of place names and geographical references will be identified by a URI (using the @ref

attribute) that will point to a formal description of the place in a formal ontology of the ancient Latin world geography (the traditional printed reference was and still is the Barrington Atlas [4]).

This ontology will be built ad hoc, reusing the data offered by the Pleiades gazetteer [5], and establishing relationships with other relevant geographic ontologies, where possible, such as Geonames. In general the ontology will be structured in a two tier fashion (following the tradition in DL ontology modelling): a T-box modelling geospatial classes of locations their properties and their relationships and an A-box with geospatial information about individual places and location. At this level the sites of antiquity will be associated with a variety of information:

- URI (and eventual links to URIS in other data sets)
- GPS coordinates
- different names, time frames of validity and etymology
- belonging to an itinerary (pilgrimage, military expedition, etc.)
- typology
- historical, geographical, cultural annotations
- links to other relevant Linked Data sets

A third level of modeling will be tied to the logical relationship between textual references (and their annotations by an encoder) and their referent in the ontology. In fact, you can easily detect that the textual context in which each geographical word (or phrase) is placed determines different modes of reference. From this point of view it seems necessary to introduce into the system an ontology of (geographic) annotations that can account for this variety of reference. In our work we will also discuss the various operational opportunities to formalize this information at the level of inline markup or through links to RDF statements in stand-off markup.

All the resources produced in our project, as he primary sources as the geographic thesaurus and the list of textual annotations that link geographic locations and places text (identified by URI) will be made available on the Web according to the principles of Linked Data, and will help to enrich the "Web of Data" with new content.

Bibliography

- [1] APA Digital Latin Library Project http://www.apaclassics.org/index.php/research/digital_latin_library_project [2] Gregory Crane call http://sites.tufts.edu/perseusupdates/2013/02/14/possible-jobs-in-digital- humanities-at-leipzig/
- [2] LASLA http://www.cipl.ulg.ac.be/Lasla/
- [3] Talbert R. (ed.), The Barrington Atlas of the Greek and Roman World, Princeton University Press 2000
- [4] Pleiades Project, http://pleiades.stoa.org/
- [5] GAP – Google Ancient Places, http://googleancientplaces.wordpress.com/
- [5] GAPvis http://googleancientplaces.wordpress.com/gapvis/
- [6] Tom Heath and Christian Bizer, Linked Data: Evolving the Web into a Global Data Space. S7nthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool 2011
- [8] Tom Elliiot, S. Gillies, "Digital Geography and Classics, in Digital Humanities Quarterly 3.1 (Winter 2009), http://www.digitalhumanities.org/dhq/vol/3/1/000031.html
- [9] Open Annotation Data Model, Open Annotation Community Group 2013, http://www.openannotation.org/spec/core/

Bringing semantic publishing in TEI: ideas and pointers

Peroni, Silvio; Vitali, Fabio

TEI has a full set of elements that can be used to describe facts about the publication details of a text, such as editionStmt, publicationStmt, and sourceDesc. A numerous list of sub-elements allows a zealous editor to provide a rich overview of publication aspects of the paper editions of the text, of this specific XML document, and of the steps through which an original source has made this XML possible. Several collections of allowable values for these elements exist, as thesauri, authority lists or simple value lists, that simplify the task to describe frequent or common situations, and that homogenize similar occurrence in different documents

of the same collection. In a way, we could characterize value thesauri as external aids to improve internal quality of digital collections of texts.

In the last few years, a new discipline has arisen, semantic publishing, that tries to improve the scientific communication by using of web and semantic web technologies to enhance a published document so as to enrich its meaning, to facilitate its automatic discovery, to enable its linking to semantically related articles, to provide access to data within the article in actionable form, and to allow integration of data between papers [1,2]. Its main interest lies in the organization and description of scientific literatures, trying to tame the incredible complexity of the modern scientific publishing environment, both in terms of size and credibility of publishing venues, authors, research groups and sponsors. For instance, SPAR [3,4,5] is a suite of orthogonal and complementary ontology modules for creating comprehensive machine-readable RDF metadata for all aspects of semantic publishing and referencing, each of them precisely and coherently covering one aspect of the publishing domain using terms with which publishers are familiar. Together, they provide the ability to describe bibliographic entities such as books and journal articles, reference citations, the organization of bibliographic records and references into bibliographies, ordered reference lists and library catalogues, the component parts of documents, and publishing roles, publishing statuses and publishing workflows. SPAR ontologies have been already used in different projects such as JISC Open Citations Project [6] – a database of biomedical literature citations, harvested from the reference lists of all open access articles in PubMed Central that reference ~20% of all PubMed Central papers (approx. 3.4 million papers), including all the highly cited papers in every biomedical field – and Semantic Web Applications in Neuromedicine (SWAN) Project [7].

One of the main aims of semantic publishing therefore is to create a rich network of interconnected facts about publications from which interesting patterns can emerge to discover, for instance, clusters of similar publications, intrinsic values of publication venues, emerging trends in publication topics, etc. In a way, we could characterize annotations coming from actual documents as internal aids to improve the external qualities of digital collections of texts, especially regarding emerging

characteristics of the collections themselves rather than belonging to individual documents.

We believe that the combination of these aspects could be mutually beneficial both in the increased quality of the individual documents, as well as in the increased quality and explorability of the emerging properties of document collections.

Being able to associate a full set of related facts to individual values in individual elements of the publication and edition details of the electronic version of a text provides the end user with a large and interesting network of considerations that go well beyond the individual text, and using standard tools from the Semantic Web may well allow reader to connect and exploit, for instance, the vast and growing collections of facts that embody the Linked Data initiative.

The actual syntax for this mesh is not particularly relevant. What is relevant is that through some syntactical mechanisms, it ends up being possible for an individual TEI document to feed Linked Data new and interesting facts about the corresponding publications and the involved actors, and conversely for Linked Data collections to enrich the amount of information about the publication and the involved actors that is made available to the interested reader, directly or after explicit queries, automatically or through the filtering and selecting action of an electronic editor.

The actual link between TEI documents and Linked Data resources is already feasible by adopting particular techniques and tools. Mainly, there are two ways to enable annotations linking existing TEI documents to Linked Data resources: either one embeds the annotation in the document itself (embedding techniques) or the annotations are stored in a separate document with references to the parts of the document each annotation refers to (standoff techniques). Neither the use embedding nor the use of standoff annotations is wrong or correct on its own; each technique has its own pros and cons that must be evaluated case by case before using them.

Even though many techniques have been devised in the past, usually the more technical solutions address only the problem of how to store the annotations, without dealing with the meaning of the annotations themself. In the case of embedded annotations, these solutions offer a generic way

to augment existing markup with annotations (e.g. RDFa [9]). In the case of standoff annotations, the existing technical solutions provide a way to address content (e.g. EARMARK [10-11] and NIF [12]). In addition to other approaches, EARMARK offers an extension [13] to actually express the meaning of the annotation and allows one to easily link bunch of text in TEI documents to external resources. It also provides a Java API [14] to support users in creating (even overlapping) annotations upon the same text, keeping track of provenance information such as the author who made the annotation and the time in which the annotation has been created. The technical solutions are only one half of what is needed to annotate documents. The other half is the use of an annotation model and vocabulary. There are many such vocabularies available, ranging from very generic annotation frameworks (e.g. the Open Annotation Data Model (OADM) [15] or the Annotation Ontology [16]), to more specific frameworks (e.g. the Linguistic Annotation Framework (LAF) [17], used to annotate the various linguistic features of a speech through its transcript, or Domeo [18], that describes annotations used to connect scholarly documents).

Bibliography

- [1] Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing, Learned Publishing 22 (2): 85–94. DOI: 10. 1087/2009202
- [2] Shotton, D., Portwin, K., Klyne, G., Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article, PLoS Computational Biology 5 (4): e1000361. DOI: 10.1371/journal.pcbi. 1000361
- [3] Semantic Publishing and Referencing Ontologies: http://purl.org/spar
- [4] Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
- [5] Peroni, S., Shotton, D., Vitali, F. (2012). Scholarly publishing and the Linked Data: describing roles, statuses, temporal and

contextual extents. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502

- [6] JISC Open Citations homepage: http://opencitations.net
- [7] Ciccarese, P., Wu, E., Kinoshita, J., Wong, G., Ocana, M., Ruttenberg, A., Clark, T. (2008). The SWAN biomedical discourse ontology, Journal of Biomedical Informatics 41 (5: 739–751. DOI: 10.1016/j.jbi.2008.04.010
- [8] Huitfeldt, C., Sperberg-McQueen, C. M. (2001). Texmecs: An experimental markup meta-language for complex documents. Working paper of the project MLCD, University of Bergen
- [9] Adida, B., Birbeck, M., McCarron, S., Herman, I. (2012). RDFa Core 1.1. W3C Recommendation, 7 June 2012. World Wide Web Consortium. http: //www.w3.org/TR/2012/REC-rdfa-core-20120607/
- [10] Di Iorio, A., Peroni, S., Vitali, F. (2011). Using Semantic Web technologies for analysis and validation of structural markup. In International Journal of Web Engineering and Technologies, 6 (4): 375-398. Olney, Buckinghamshire, UK: Inderscience Publisher. DOI: 10.1504/IJWET.2011.043439
- [11] Di Iorio, A., Peroni, S., Vitali, F. (2011). A Semantic Web Approach To Everyday Overlapping Markup. In Journal of the American Society for Information Science and Technology, 62 (9): 1696-1716. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. DOI: 10.1002/asi.21591
- [12] Hellmann, S., Lehmann, J., Auer, S. (2012). Linked-data aware uri schemes for referencing text fragments. In ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Aquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (Eds.), Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Lecture Notes in Computer Science 7603: 398-412. Berlin, Germany: Springer. DOI: 10.1007/978-3-642-33876-2_17
- [13] Peroni, S., Gangemi, A., Vitali, F. (2011). Dealing with Markup Semantics. In Ghidini, C., Ngonga Ngomo, A.,

Lindstaedt, S. N., Pellegrini, T. (Eds.), Proceedings the 7th International Conference on Semantic Systems (I-SEMANTICS 2011): 111-118. New York, New York, USA: ACM. DOI: 10.1145/2063518.2063533

- [14] Barabucci, G., Di Iorio, A., Peroni, S., Poggi, F., Vitali, F (2013). Annotations with EARMARK in practice: a fairy tale. Submitted for publication in the 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE 2013)
- [15] Sanderson, R., Ciccarese, P., de Sompel, H. V. (2013). Open annotation data model. W3C Community draft, 08 February 2013. http: //www.openannotation.org/spec/core/20130208/
- [16] Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T. (2011). An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics, 2 (2): 1–24. DOI: 10.1186/2041-1480-2-S2-S4
- [17] ISO (2012). ISO 24612:2012 Language resource management — Linguistic annotation framework (LAF). ISO
- [18] Ciccarese, P., Ocana, M., Clark, T. (2012). Open semantic annotation of scientific publications using DOMEO. Journal of Biomedical Semantics, 3 (1): 1–14. DOI: 10.1186/2041-1480-3-S1-S1

Posters

Library of components for the Computational Philological Domain dealing with TEI markup guidelines CoPhiLib

Boschetti, Federico; Bozzi, Andrea; Del Grosso, Angelo Mario

The aim of this poster is to illustrate the Collaborative Philology Library (CoPhiLib), a library of components devoted to editing, visualizing and processing TEI annotated documents in the subdomain of philological studies. The overall architecture is based on the well known Model-View-Controller (MVC) pattern, which separates the representation of data from the rendering and management (business logic) of the content, for the sake of flexibility and reusability. The CoPhiLib library maps the annotated document on an aggregation of objects, visualized via web as a collection of widgets rendered on the client through rich standard web technologies such as html5, css3, jquery, ajax etc. and controlled by special components devoted to monitor the behavior and interactions among the other components.

The specifications, expressed using the Unified Modelling Language (UML), are language independent and stay at the top level of abstraction, as a formal guidelines for the actual implementations, for example using the Java programming language or any other which follows the object oriented programming paradigm as it could be Python. Currently, only a very small subset of TEI tags are taken into account in our specifications, because our approach is a trade-off between a top-down and a bottom-up design. The approach is top-down, because we analyze the high-level behaviors of the interacting objects and the use-case with related scenarios among functionalities that agents are expected to use. But it is also bottom-up, because we develop applications for specific projects, such as Greek into Arabic or Saussure Project, and we refactor the original design of the specific projects when upper levels of abstraction, valid for multiple scenarios, can be identified and the new interfaces must be taken into account in order to update and extend the basic functionalities.

According to the specifications, the APIs and the actual libraries are developed. The current implementation of the CoPhiLib library is

based on the Java platform and the overall system has been developed following the Java enterprise powerful programming model Server Faces Framework (JSF2). Documents are stored in a database XML oriented: eXist-db, but different cross-platform solutions can be easily adopted by implementing a data access object (DAO pattern), due to the pluggable structure. Our application is designed as a collaborative multilayer application and handles the presentation logic by making use of the world wide web (web) Java technologies and the best practices like facelets templates to minimize code and maximize reuse as well as a complete rich Ajax composite component taglib, in order to offer a friendly and efficient web graphical user interface (the most popular is RichFaces alongside to IceFaces, but we preferred PrimeFaces as the most rising one). In the field of digital scholarship users mainly ask web applications that allow an easy access to resources (both textual and visual) and that provides the possibility to work in a collaborative environment by comparing resources, creating relations among resources, adding notes and comments or critical apparatus and sharing them.

From the collection of the TEI-compliant documents stored for the specific projects, the scheme is read (or dynamically generated and read). The actual scheme is expected to be a small subset of the TEI schemes (as discussed above) and it is used by the applications developed with the COPhiLib, in order to instruct the factories on how to instantiate the objects that implement or extend the interfaces or the abstract classes.

This structure provides the necessary flexibility to adapt, at run time, the same application to different uses, according to the nature of the chunks of information contained in the documents that must be rendered. For example, the abstract model is able to manage different multimedia resources in parallel for scholarly editions, like in the E.R.C. Greek into Arabic project, and it is able to deal with facsimile manuscript images within the related transcription, like in the P.R.I.N. Saussure Edition project or, in the future, to provide a sheet music viewer with the related midi or wave execution. Different instances of the Model are obtained by serializing the TEI document through a marshall and unmarshall process, obtaining a synchronized and uniform state of the stored data.

CoPhiLib handles textual phenomena by separating the structure of the text (codicological aspects) from its analyses (philological, linguistic, metric, stylistic, etc.). Stand-off markup approach has been used to manage the data arisen from the automatic text analysis.

Bibliography

- Bozzi, Andrea 'G2A: a Web application to study, annotate and scholarly edit ancient texts and their aligned translations' Stuida graeco-arabica Pacini Editore Pisa 2013
- Burbeck, Steve *Applications Programming in Smalltalk-80TM: How to use Model-View-Controller (MVC)* 1992 [35]
- Del Grosso, Angelo Mario Boschetti, Federico 'Collaborative Multimedia Platform for Computational Philology, CoPhi Architecture' *IARIA Conference: proceedings of MMEDIA 2013, The Fifth International Conference on Advances in Multimedia* Venice 2013
- Fowler, Martin *Analysis Patterns: Reusable Object Models* Addison-Wesley Menlo Park, Calif. ; Harlow 1996
- Gamma, Erich Helm, Richard Johnson, Ralph Vlissides, John *Design Patterns: Elements of Reusable Object-Oriented Software* Addison-Wesley Boston, MA, USA 1995
- Hohpe, Gregor Woolf, Bobby *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions* Addison-Wesley Boston, MA, USA 2004
- Burnard, Lou Bauman, Syd *TEI P5: Guidelines for Electronic Text Encoding and Interchange* Oxford 2008 [36]

[35] http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html

[36] . http://www.tei-c.org/Guidelines/P5

TEI as an archival format

Burnard, Lou; Larousse, Nicolas

The adoption of the TEI as a common storage format for digital resources in the Humanities has many consequences for those wishing to interchange, integrate, or process such resources. The TEI community is highly divers, but there is a general feeling that all of its members share an understanding of the best way to use the TEI Guidelines, and that those Guidelines express a common understanding of how text formats should be documented and defined. There is also (usually) a general willingness to make resources encoded according to the TEI Guidelines available in that format, as well as in whatever other publishing or distribution format has been adopted by the project. The question arises as whether such TEI-encoded resources are also suitable for long term preservation purposes : more specifically, if a project wishes to ensure long term preservation of its resources, should it archive them in a TEI format? And if so, what other components (schema files, stylesheets, etc.) should accompany the primary resource files when submitting them for long term preservation in a digital archive? TEI encoded resources typically contain mostly XML-encoded text, possibly with links to files expressed using other commonly encountered web formats for graphics or audio; is there any advantage to be gained in treating them any differently from any other such XML encoded resource?

This is not an entirely theoretical question : as more and more digitization projects seek to go beyond simply archiving digital page images, the quantity of richly encoded TEI XML resources representing primary print or manuscript sources continues to increase. In France alone, we may cite projects such as the ATILF, OpenEditions, BVH, BFM, Obvil and many more for all of which the TEI format is likely to be seen as the basic storage format, enabling the project to represent a usefully organised structural representation of the texts, either to complement the digital page images, or even to replace them for such purposes as the production of open online editions. When such resources are deposited in a digital archive, how should the archivist ensure that they are valid TEI and will continue to be usable ? One possibility might be to require that such

resources are first converted to some other commonly recognised display format such as PDF or XHTML; and indeed for projects where the TEI form is considered only as a means to the end of displaying the texts, this may well be adequate. But since TEI to HTML or TEI to PDF are lossy transformations, in which the added value constituted by TEI structural annotation is systematically removed this seems to us in general a less than desirable solution. We would like to be able to preserve our digital resources without loss of information, so as to facilitate future use of that information by means of technologies not yet in existence. Such data-independence was, after all, one of the promises XML (and before it SGML) offered.

The data archivist needs to be able to test the coherence and correctness of the resources entering the archive, and also to monitor their continued usability. For an XML-based format, this is a relatively simple exercise. An XML file must be expressed using one of a small number of standard character encodings, and must use a tagging system the syntactic rules of which can be written on the back of a not particularly large envelope. The algorithm by which an XML document can be shown to be syntactically correct, (”well formed”) is expressible within the same scope and producing a piece of software able to determe that correctness is consequently equally trivial. The XML Recommendation adds a layer of "syntactic validation" to this, according to which the use of XML tags within a set of documents can be strictly controlled by means of an additional document known as a schema, defining for example the names of all permitted XML elements and attributes, together with contextual rules about their valid deployment. Syntactic validation of an XML resource against its schema is also a comparatively simple and automatic procedure, rerquiring only access to the schema and an appropriate piece of software. (Given the dominant position enjoyed by XML as a data format, the current wide availability of reliable open-source validators for it seems unlikely to change, even in the long term)

However, the notion of "TEI Conformance" as it is defined in the current Guidelines goes considerably beyond the simple notion of syntactic validity. An archivist concerned to ensure the coherence and correctness of a new resource at this further level needs several additional tools and

procedures, and a goal of our project is to determine to what extent the goal of ensuring such conformance is quixotic or impractical. In particular, we will investigate the usefulness of the TEI's ODD documentation format as a means of extending the scope of what is possible in this respect when using a conventional XML schema language such as RELAX NG or ISO Schematron.

Our initial recommended approach for ingest of a conformant TEI resource might include :

- syntactic validation of each document against the most appropriate TEI schema; for documents containing textual data this would naturally include TEI All, but also any project-supplied XML schema, and also (for any ODD document supplied) the standard TEI ODD schema;
- creation of a TEI schema from the supplied ODD and validation of the documents against that in order to validate any project-specific constraints such as attribute values;
- comparison of the ODD supplied with an ODD generated automatically from the document set;
- definition and usage of a set of stylesheets to convert the resource into a "lowest common denominator" TEI format

Such an approach suggests that the "submission information package" for a TEI resource will contain a number of ancillary documents or references to documents, notably to a TEI P5-conformant ODD from which a tailored set of syntactic and semantic validators can be generated using standard transformations. We hope to report on this and on the results of our initial experiments with some major French-language resources at the Conference.

The Open Bibliography Project

Childress, Dawn; Clair, Kevin

Humanities scholars often create bibliographies in the course of their work. These can take on many forms: annotated bibliographies, descriptive bibliography and catalogues, author and subject bibliographies, or learning objects for scholars researching people and concepts in their field. The aggregate nature of these publications means that printed bibliographies are often outdated soon after publication and calls for a shift away from print to a more dynamic, web-based bibliography that allows updating and revising as new information becomes available.

While many bibliographical works are still published as print monographs, web-based bibliographies are nothing new; however, current web-based bibliography publishing models present a number of challenges to those wanting to share their research openly on the web. The creators of scholarly web bibliographies must design, create, and host relational databases, forms, queries, and a web interface, as well as deal with the hosting, access and maintenance issues associated with publishing a searchable, accessible database to the web. Most humanities scholars and librarians do not have the technological skills nor access to the infrastructure necessary to host such a site and libraries and institutions are not always able to accommodate these "boutique" project requests.

Additionally, these bibliographies are often multi-layered documents, rich with bibliographic information, metadata about the items described, and added value in the form of annotations, contextual information, and links to other relevant information and resources. This bibliographic and contextual information, which in many cases cannot be found anywhere else on the Web, would be extremely valuable to other researchers if made available in a data markup format that is open to harvesting and repurposing. Scholars working on publishing their own bibliographies would also benefit from an automated approach to harvesting and aggregating bibliographic information into their own bibliographies and publishing that information using open standards.

The Open Bibliography Project [1] represents a novel approach for publishing bibliographies to the Web using TEI in a format that enables linking, sharing, and repurposing of the content of TEI-encoded scholarly bibliographies. To that end, the project has two goals: a) to develop tools allowing scholars to easily construct, markup, and publish bibliographies in more meaningful ways while exposing their structured data to other Web applications; and b) to build a vocabulary for marking up and transforming structured bibliographic data within these documents, using existing vocabularies such as TEI and schema.org to the extent possible, and creating new terms where necessary. Ultimately we would like to provide a tool for scholars to construct bibliographies, assigning structure to citations and annotations using a Web form (XForms or similar technology), and providing a mapping for linking to occur in the background.

The Project is built around a custom TEI module for describing multiple types of bibliographies, including annotated bibliographies and descriptive bibliography, with XSL and CSS stylesheets for transforming the TEI-encoded documents into searchable, structured web (or print) editions and possibly into interactive maps and data visualizations. Using a custom TEI module with pre-defined stylesheets means a lightweight, low-barrier publishing solution for researchers that requires only minimal knowledge of XML and basic web hosting, such as a web folder on a university server or Google Sites.

The Project recognizes the need for sharing unique bibliographic and contextual information found in bibliographies with the wider web of scholarly data in the humanities, social sciences, and other disciplines. Using linked open data standards, such as microdata, the Project hopes to further enhancing the value of scholarly bibliographies by linking them to the linked open data web. Because they are highly structured documents, TEI bibliographies easily lend themselves to linked open data markup; in addition, the annotations within them provide context about items contained within them that may not exist elsewhere on the Web. Initiatives such as schema.org [2] provide tools for document markup compatible with the linked data cloud, and projects such as VIVO [3]

provide examples of how faculty profiles and CVs, published as structured bibliographic data, may be published electronically.
Defining a proof of concept for the idea is the first stage of this project. Using the Three Percent translation database, published by the University of Rochester [4], as our seed data, we intend to demonstrate how TEI-encoded bibliographic metadata may be published as linked data in a variety of markup formats and included in the linked data ecosystem. We plan to develop a simple vocabulary for marking up individual citations in the database with schema.org attributes, to which we may map the Three Percent database elements. We will share our XSLT stylesheets under an open license on the Web, so that interested scholars and researchers may contribute to its continued development.

Bibliography

- [1] http://dawnchildress.com/obp
- [2] http://schema.org
- [3] http://vivo.library.cornell.edu
- [4]http://www.rochester.edu/College/translation/threepercent

An easy tool for editing manuscripts with TEI

Dumont, Stefan; Fechner, Martin

The Berlin Brandenburg Academy of Sciences and Humanities (BBAW) is home to multiple long term research projects which encompass various fields of study. The research group TELOTA (The Electronic Life of the Academy) supports the digital humanities aspects of these projects, including developing software solutions for the daily work of their researchers.
Experience shows that the readiness to use TEI encoding for the digital transcription and annotation of manuscripts greatly relies on the user-

friendliness of the entry interface. From the perspective of a researcher, working directly in XML is a backwards step in comparison to programs like MS Word. A new software solution must therefore at least offer the same amount of editorial comfort as such programs. Ideally, it would also encompass the complete life-cycle of an edition: from the first phases of transcription to the final publication.

Last year TELOTA developed such a software solution for the recently begun scholarly edition project Schleiermacher in Berlin 1808-1834. The solution consists of various software components that allow the researchers to construct and edit transcriptions of Schleiermachers manuscripts into XML following the TEI guidelines. It includes the possibility to create apparatuses of different kinds, as well as to createwithout much additional effortboth a print and web publication.

The new digital Schleiermacher edition is based on XML schemata, written according to the guidelines of the TEI. A TEI schema was created for each manuscript type: letters, lectures, and a daily calendar. The three schemata however all share a core group of elements. All text phenomena as well as editorial annotations are represented through TEI elements and attributes. The schemata were formed from the sections of the TEI guidelines which suited the projects needs. The addition of project-unique elements or attributes was unnecessary.

The central software component of the new digital work environment is Oxygen XML Author. The researcher does not edit the XML code directly, but instead works in a user-friendly Author mode, which is designed through Cascading Stylesheets (CSS). The researcher is able to choose more than one perspective within the Author view, and thus can select per mouse click the appropriate perspective for the current task. Additionally, a toolbar is provided with which the researcher can enter markup with the push of a button. In this way text phenomena such as deletions or additions, or editorial commentary, are easily inserted. Person and place names can also be recorded with their appropriate TEI markup, and in addition they can be simultaneously linked to the corresponding index. This is done through selecting the name from a convenient drop down list. The entire manuscript text can thus be quickly and simply marked up with TEI conform XML.

Besides creating a digital work environment in Oxygen XML Author, a website was also built for the project based on eXist, XQuery, and XSLT. Through the website the researchers can easily page through or search the current data inventory. For instance, letters can be filtered through correspondence partner and/or year. The user can also follow a correspondence series according to the selected person, or find all texts in which a person was mentioned. The website is presently only available for the project staff, but it offers a prototype for the future, publicly accessible, website.

With the help of ConTeXt a further publication type, a print edition, is automatically generated as a PDF from the TEI XML document. The layout and format is based on the previously printed volumes of the critical edition for Friedrich Schleiermachers works. Each TEI element is given a specific formatting command through a configuration file. In this way the different apparatuses appear as footnotes that refer to the main text with the help of line numbers and lemmata. The print edition can also provide the suitable index for each transcription and solves any occurring cross references between manuscripts.

This work environment has been in use for a year by the research staff of the Schleiermacher edition for their daily work. When asked their opinion, the researchers offered predominantly positive feedback. The only criticism was the fact that the text became difficult to read when it included a great deal of markup. TELOTA met this concern by adding more Cascading Stylesheets, thus allowing for different views of the text that showed only specific groups of elements. The researchers were however in absolute agreement that the new work environment greatly eased their editorial work and saved them significant time. The possibility to directly check the results of their work in a web presentation or as a printed edition was seen as very positive. Such features let the user experience per click the advantages of encoding with TEI. The staff also expressed their relief that it was unnecessary to work directly in XML, and that they instead could mark up their texts through a graphic and easy to use interface.

After the success of the pilot version, the work environment will be implemented this year for further academy projects. The TEI XML

schemata and main functions that make up the basis of the work environment can be customized to the different manuscript types and its needs. Furthermore, this solution has already been adapted by other institutions, such as the Academy of Sciences and Literature in Mainz.

Bibliography

- Dumont, Stefan; Fechner, Martin: Digitale Arbeitsumgebung für das Editionsvorhaben »Schleiermacher in Berlin 1808—1834« In: digiversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften. URL: http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834
- Burnard, Lou; Bauman, Syd (Hg.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Charlottesville, Virginia, USA 2007. URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
- User Manual Oxygen XML Author 14. URL: http://www.oxygenxml.com/doc/ug-editor/
- eXist Main Documentation. URL: http://www.exist-db.org/exist/documentation.xml
- ConTeXt Dokumentation. URL: http://wiki.contextgarden.net/Main_Page

eCodicology - Algorithms for the Automatic Tagging of Medieval Manuscripts

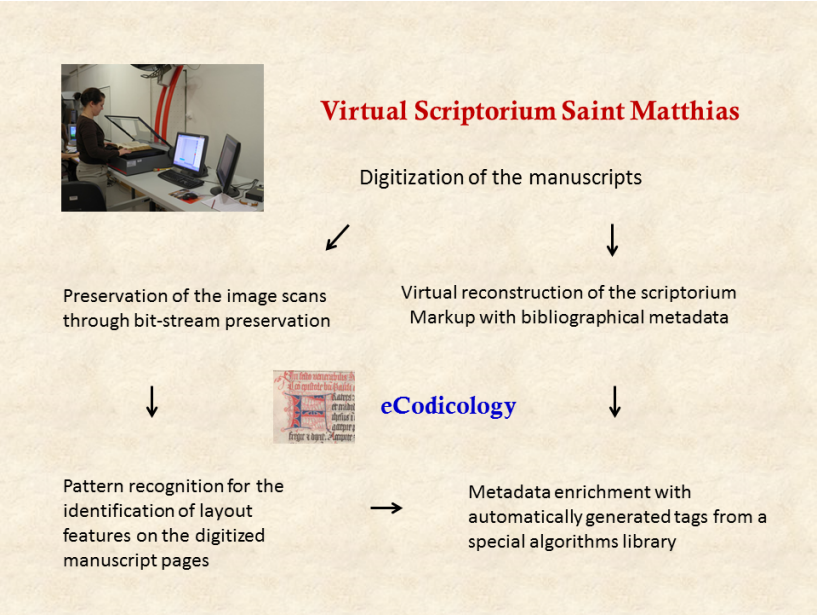
Embach, Michael; Krause, Celia; Moulin, Claudine; Rapp, Andrea; Rindone, Francesca; Stotzka, Rainer; Tonne, Danah; Vanscheidt, Philipp

General description

eCodicology uses the library stock of roughly 500 medieval manuscripts which have been collected in the Benedictine Abbey of St. Matthias in Trier (Germany). The manuscripts were digitized and enriched with bibliographic metadata within the scope of the project Virtuelles Skriptorium St. Matthias / Virtual Scriptorium Saint Matthias (http://stmatthias.uni-trier.de/http://stmatthias.uni-trier.de/). Funded by the German Research Foundation (DFG), digital copies were created in the city library of Trier, the long term preservation is undertaken at the University of Trier. The purpose of the BMBF-funded project eCodicology is the development, testing and optimization of new algorithms for the identification of macro- and microstructural layout elements on these manuscript pages in order to enrich their metadata in XML format according to TEI standards.

The database of the St. Matthias project holds basic information on the physical properties of manuscripts, as they have been described in the older manuscript catalogues. Essential components of each manuscript description are details of the layout features of the manuscript. These details are in part fragmentary and incomplete and can therefore be refined and completed by means of automatic tagging. The more precisely and elaborately those details are described, the better comparisons and analyses can be performed. Within the scope of eCodicology, the first step is the creation of an inventory of features defining those elements to be recognized reliably with the aid of the algorithms for feature extraction. On this basis, it is expected that new scientific findings about corpora of writers, writing schools, references between manuscripts and proveniences become possible. The great amount of image scans will be analyzed on an empirical basis with the aim that the subjective view of the codicologist can - as it were - get objectified.

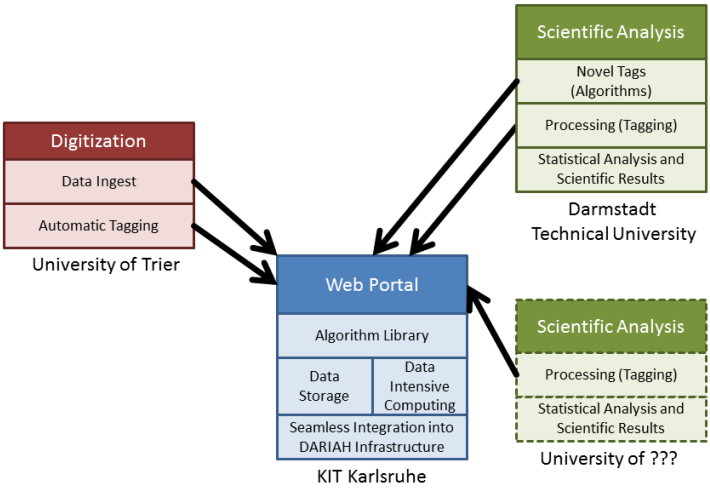
As can be seen from the figure below, the data that has been produced in the project Virtuelles Skriptorium St. Matthias is the starting point of the work in eCodicology. The image scans are hosted on distributed servers and are synchronized regularly. Based on this initial data the previous catalogues can be automatically enriched and refined by use of feature extraction processes.



Aims of the project partners and technical procedure

The eCodicology project is managed by three project partners working on different tasks (see the figure below). The digitized images are processed at the Karlsruhe Institute of Technology (KIT) using a library consisting of image processing and feature extraction algorithms which are defined in close collaboration between humanities scholars and computer scientists. The metadata schema for the processing and the models for the XML files, in which the results will be saved, are developed in Trier as well as in Darmstadt on the basis of TEI P5. The scientific evaluation will finally take place in Darmstadt. Additionally, statistical analysis of the

manuscript groups will be performed. It shall be possible to conduct, adapt or extend the scientific evaluation at any other university.



A software framework will automate the procedure of complex data analysis workflows and is designed generically so that a great amount of image data can be processed with any desired algorithm for feature extraction (basic components: ImageJ and MOA/Weka). Since it will be adaptable for a wider range of documents, the framework will be integrated as a service into the DARIAH infrastructure (http://de.dariah.eu/http://de.dariah.eu/). The algorithm library is implemented specifically for the automatic analysis of medieval manuscripts. New algorithms can be created by the users at any time and they can be integrated into the library through the web portal. The configuration of the processes, the selection of the algorithms for feature extraction from the algorithm library and their parameterization are controlled via the web portal.

Processing and metadata schema

The processing of a codex page normally entails the following steps:

1 Preparation and normalization of the page: this contains basic image processing steps such as e.g. the alignment of the page, white balance, histogram operations for the normalization of contrasts.

2 Object segmentation: the segmentation separates image objects (e.g. writing and illustrations) from the background. The complexity of this process can vary and it is one of the most elaborate operations in the digital image processing.

3 Feature extraction: features describing the whole page and the segmented objects can be measured using the algorithm library.

4 Storage: the extracted features are stored within the metadata of the codex image.

The metadata schema used in the DFG-project Virtuelles Skriptorium St. Matthias corresponds to the METS format, as it is used for the DFG Viewer (http://dfg-viewer.de/en/regarding-the-project/http://dfg-viewer.de/en/regarding-the-project/). Instead of MODS a TEI header is used, which is more specifically adapted to the demands of a manuscript description. A refining of the metadata is intended especially for the measurements of the following basic layout features: page dimensions (height and width), text or writing space, space with pictorial or graphical elements, space with marginal notes and glosses. Additionally, the absolute number of lines, headings, graphical or coloured initial letters or rubricated words and sentences will be annotated. It is also intended to find a way to tag the position of graphical elements or text blocks on each page. From these data certain relations or proportions can be deduced. These relations may tell us for example something about special patterns or layout types.

At the moment, the refining of data concentrates on the elements objectDesc with supportDesc and layoutDesc as well as decoDesc. A focus is laid especially on the following fields (the respective TEI tag is set in brackets):

1 Layout information (layout): conceivable attributes are @ruledLines, @writtenLines and @columns. Also, the type area or page design can be measured exactly.
2 Dimensions (dimensions, extent, height, width): attributes allow also a TEI-compliant description of minimum, maximum and average information (@atLeast, @atMost, @min, @max).
3 Information on perceivable visual units like initials, marginal decoration and embedded images (decoNote, @initial, @miniature, @border), rubrications, additional notes and eventually also multiple foliations.

A first draft of the metadata schema can give a short glimpse on some adaptations concerning the physical description of manuscripts:

```
<physDesc><objectDesc form="codex"><supportDesc>
. . .
<extent>
<measure unit="leaves" quantity="100"></measure>
<locusGrp xml:id="locusGrp001"><locus from="1" to="100"></locus></locusGrp>
<measureGrp type="leaves" corresp="#locusGrp001">
<height quantity="250" unit="mm">250mm</height>
<width quantity="150" unit="mm">150mm</width>
</measureGrp>
<measureGrp type="binding">
<height quantity="275" unit="mm">275mm</height>
<width quantity="175" unit="mm">175mm</width>
<measure type="spineHeight">4°</measure>
</measureGrp>
</extent>
. . .
</supportDesc><layoutDesc><layout columns="2" writtenLines="24">
<locusGrp>
<locus from="1" to="100" xml:id="locusGrp002"></locus>
</locusGrp>
<dimensions type="written" corresp="#locusGrp002">
<height quantity="200" unit="mm" min="199" max="201" confidence="0.8">
200mm
</height>
<width quantity="100" unit="mm" min="98" max="101" confidence="0.75">
100mm
</width>
</dimensions>
</layout><layout ruledLines="32">
<locusGrp>
<locus from="1r" to="202v" xml:id="locusGrp003"></locus>
</locusGrp>
</layout></layoutDesc></objectDesc>
. . .
<decoDesc>
<decoNote type="initial"></decoNote>
<decoNote type="miniature">
```

```
<locusGrp>
<locus>8v</locus></locusGrp>
<dimensions>
<height quantity="50" unit="mm" min="49" max="51" confidence="0.8">50mm
</height>
<width quantity="50" unit="mm" min="49" max="51" confidence="0.8">50mm
</width>
</dimensions>
</decoNote>
<decoNote type="border"></decoNote>
</decoDesc>
</physDesc>
```

Based on the exemplary interpretation of the empirical data the sustainability of the approach as well as the validity of the inventory of layout features have to be proven. The drawing up of sophisticated microscopic information and metrics on every single manuscript page subsequently allows an evaluation of the codices from the abbey of St. Matthias on the basis of quantitative methods: hereby, tendencies throughout the times related to certain genres or languages can be described in a highly elaborated way, image-text-proportions (text space vs. image space) can be defined exactly and relationships to epochs, genres, contents and functions can be created.

Bibliography

- Embach, Michael; Moulin, Claudine (Ed.): Die Bibliothek der Abtei St. Matthias in Trier - von der mittelalterlichen Schreibstube zum virtuellen Skriptorium, Trier 2013.
- Tonne, Danah; Rybicki, Jedrzej; Funk, Stefan E.; Gietz, Peter: Access to the DARIAH Bit Preservation Service for Humanities Research Data, in: P. Kilpatrick; P. Milligan; R. Stotzka (Ed.), Proceedings of the 21th International Euromicro Conference on Parallel, Distributed, and Network-Based Processing, Los Alamitos 2013, pp. 9-15.
- Tonne, Danah; Stotzka, Rainer; Jejkal, Thomas; Hartmann, Volker; Pasic, Halil; Rapp, Andrea; Vanscheidt, Philipp; Neumair, Bernhard; Streit, Achim; García, Ariel; Kurzawe, Daniel; Kálmán, Tibor; Rybicki, Jedrzej; Sanchez Bribian, Beatriz: A Federated Data Zone for the Arts and Humanities, in: R. Stotzka; M. Schiffers; Y. Cotronis (Ed.), Proceedings of the 20th International Euromicro Conference on Parallel, Distributed, and Network-Based Processing, Los Alamitos 2012, pp. 198-205.

• Vanscheidt, Philipp; Rapp, Andrea; Tonne, Danah: Storage Infrastructure of the Virtual Scriptorium St. Matthias, in: J. C. Meister (Ed.), Digital Humanities 2012, Hamburg 2012, pp. 529-532.

ReMetCa: a TEI based digital repertory on Medieval Spanish poetry

González-Blanco García, Elena; Rodríguez, José Luis

The aim of this talk is to present a Digital Humanities-TEI project devoted to create a computer-based metrical repertory on Medieval Castilian poetry (ReMetCa,www.uned.es/remetca). It will gather poetic testimonies from the very beginnings of Spanish lyrics at the end of 12th century, until the rich and varied poetic manifestations from the Cancioneros of the 15th and 16th centuries. Although metrical studies on Spanish Medieval poetry are developing fast in the last years, researchers have not created a digital tool yet, which enables to undertake complex analysis on this corpus, as it has already been done in other lyrical traditions in Romance languages, such as the Galician-Portuguese, Catalan, Italian or Provençal lyrics, among others, where the first digital repertories arose. ReMetCa is conceived as an essential tool to complete this digital poetic puzzle, which will enable users to develop powerful searches in many fields at the same time, thanks to the possibilities offered by new technologies. It will be very useful for metrical, poetic and comparative studies, as well as a benchmark to be linked to other international digital repertories.

This project is based on the integration of traditional metrical and poetic knowledge (rhythm and rhyme patterns) with Digital Humanities technology: the TEI-XML Markup Language and his integration in a Relational Database Management System which opens the possibility to

undertake simultaneous searches and queries using a simple searchable user-friendly interface.

Starting point: poetic repertories in European lyrics

Three significant periods can be distinguished in the creation of medieval and renaissance poetic repertoires. The first one matches up with Positivism (end of the 19th century), with the works of Gaston Raynaud (1884), Gotthold Naetebus (1891), and Pillet and Carstens (1933), among others. The second one starts after the Second World War with the classic work of Frank on Provencal troubadours' poetry (1953-57), and continues during long time with the editions of printed metrical repertoires (in Old French lyrics Mölk and Wolfzettel, in Italian Solimena, Antonelli, Solimena again, Zenari, Pagnotta, and Gorni, in the Hispanic philology Tavani, Parramon i Blasco, and Gómez Bravo, in the German Touber and the *Repertorium der Sangsprüche und Meisterlieder*.

Technological advances have made it possible to create a third generation of repertoires –made and searchable with a computer– in which time of research is considerably reduced. The first digital poetical repertoire was the *RPHA* (*Répertoire de la Poésie hongroise ancienne jusqu'à 1600*) published by Iván Horváth and his group in 1991. Galician researchers created *Base de datos da Lírica profana galego-portuguesa (MedDB)*; Italian researchers digitalized *BEdT* (*Bibliografia Elettronica dei Trovatori*); later appeared the *Nouveau Naetebus*, the *Oxford Cantigas de Santa María Database*, the *Analecta Hymnica Digitalia*, etc.

All these repertoires are very valuable, as they enhance the possibilities of performing comparative researches. The Spanish panorama looks, however, weak in this area, as we do not have a poetic repertoire which gathers the metrical patterns of Medieval Castilian poetry (except for the book of Ana María Gómez Bravo (1999), restricted to Cancionero poetry).

Researchers are, however, more and more conscious of the importance of metrical studies to analyze and to understand Spanish Medieval poetry, as it has been recently shown by the bibliographic compilations of José María Micó (2009) or Vicenç Beltrán (2007). On the other hand, metrical studies have flourished thanks to the creation of specialized journals, such as *Rhythmica. Revista española de métrica comparada*, edited by

Universidad de Sevilla (ISSN 1696-5744), created in 2003 and directed by Domínguez Caparrós and Esteban Torre, or the *Stilistica e metrica italiana* (2001) directed by Pier Vincenzo Mengaldo, as well as the digital journal *Ars Metrica* (*www.arsmetrica.eu* ISSN 2220-8402), whose scientific committee is composed by researchers from different countries.

Other important focuses of recent metrical studies have been research projects, whose results are being published as articles in books and journals and also as PhD works and thesis. There have also been organized several meetings and seminars concerning metrical and poetic problems. In this sense, it is worth to mention the project of prof. José Domínguez Caparrós on metrics in the 20th century, and the one leaded by prof. Fernando Gómez Redondo, devoted to write a diachronic history on medieval Castilian metrics by using traditional definitions of vernacular metrics.

As far as the integration of philology and computer technology is concerned, there have been significant advances during the last years in Spain (it is worth to mention some projects like Beta-Philobiblon *http://bancroft.berkeley.edu/philobiblon/beta_es.html*), or the digital editions of Lemir (*http://parnaseo.uv.es/lemir.htm*), or the digital bulletin of the AHLM (*www.ahlm.es*), as well as the upgrades and improvements made by the Biblioteca Virtual Cervantes (*http://www.cervantesvirtual.com/*). These tools show, however, a lack of metrical analysis of the texts and do not usually offer any metrical information about them, and this is the aspect that we want to improve with our tool ReMetCa.

Specific goals of this project and tool:

With the creation of ReMetCa our main goals are:

- To create a database that integrates the whole known Castilian poetic corpus from its origins up to 1511 (over 10.000 texts).
- To systematize traditional metric analysis by creating different tags suitable for all the poems of the corpus.
- To provide access to metrical schemes altogether with texts, as well as data sheets gathering the main philological aspects that characterize the poems.

- To develop a TEI-based description and make it available to all research community through a Web Application based on a Relational Database Management System and PHP.
- To follow the Standards for Web Content Interoperability through Metadata exchange that will allow the future integration of our project in a megarepertoire, the *Megarep* project, in which Levente Seláf (González-Blanco and Seláf 2013), a Hungarian researcher of ELTE University, is already working.
- To contribute to the improvement and discussion about TEI, specifically the TEI-Verse module.
- To promote research in Digital Humanities within the area of Philology, Metrics, and Literary Theory in Spain.

Technical issues

This poster will be focused on the set of elements of TEI-Verse's module. Every element will be represented with UML as an entity with its attributes and relationships. The result of this representation will be a complete conceptual model, which will work as the starting point of the logical model, build with an *Entity-Relationship (ER)* diagram.

The next step is the creation of the physical model, and it will provide us with the opportunity to discuss on the appropriateness of a Relational *Database Management System*, compared to the apparently easier option of using a native database XML. We will consider pragmatic aspects, such as the usual familiarity of most web applications programmers with RDBMS and the possibility of combining instances of relational systems with documents XML.

The choice of a concrete RDBMS will present two possibilities: MySQL with XPath or Oracle, with its columns XMLType and its incorporation to the recent versions (10g Release 2) of the XQuery query language. Both models, conceptual and logical, will be implemented in both RDBMS fully developed.

A series of queries SQL will be launched on this operative installation, especially centered on data extraction with XPath, in order to verify the actual behavior of each proposal. To perform this simulation, we will use our actual project records and we will simulate obtaining useful data for

research that could have been proposed as a requisite of this application by a researcher specialist in the field.
To finish, we will propose a web application with forms for data introduction made with a PHP framework, such as CodeIgniter.
We would like to present these solutions in this poster to be able to discuss them with the TEI community and with members of other projects working with TEI-verse.

REFERENCE WORKS

Repertoires and digital databases

- Répertoire de la poésie hongroise ancienne, (Iván Horváth *et alii*) *http://magyar-irodalom.elte.hu/repertorium/*, *http://tesuji.eu/rpha/search/rpha5*
- MeDBD — Base de datos da Lírica profana galego-portuguesa, (Mercedes Brea *et alii*) *http://www.cirp.es/bdo/med/meddb.html*
- BEdT — Bibliografia Elettronica dei Trovatori, (Stefano Asperti, Fabio Zinelli *et alii*) *www.bedt.it*
- Dutch Song Database (Louis Grijp *et alii*): *http://www.liederenbank.nl/index.php?lan=en*
- The Oxford Cantigas de Santa Maria Database (Stephen Parkinson) *http://csm.mml.ox.ac.uk/*
- Le Nouveau Naetebus — Répertoire des poèmes strophiques non-lyriques en langue française d'avant 1400 (Levente Seláf) *nouveaunaetebus.elte.hu*
- Analecta Hymnica Medii Aevi Digitalia, (Erwin Rauner), *http://webserver.erwin-rauner.de/crophius/Analecta_conspectus.htm*

Metrical repertoires published in paper

- Antonelli, R., *Repertorio metrico della scuola poetica siciliana*, Palermo, Centro di Studi Filologici e Linguistici Siciliani, 1984.
- Betti, Maria Pia, *Repertorio Metrico delle Cantigas de Santa Maria di Alfonso X di Castiglia*, Pisa, Pacini, 2005.
- Brunner, Horst, Burghart Wachinger et Eva Klesatschke, *Repertorium der Sangsprüche und Meisterlieder des 12. bis 18. Jahrhunderts*, Tubingen, Niemeyer, 1986-2007.

- Frank, Istvan, *Répertoire métrique de la poésie des troubadours*, Paris, H. Champion, 1966 [Bibliotheque de l'Ecole des hautes etudes. Sciences historiques et philologiques 302, 308].
- Gorni, Guglielmo, *Repertorio metrico della canzone italiana dalle origini al Cinquecento (REMCI)*, Florencia, Franco Cesati, 2008.
- Gómez Bravo, Ana María, *Repertorio métrico de la poesía cancioneril del siglo XV*, Universidad de Alcalá de Henares, 1999.
- Mölk, Ulrich y Wolfzettel, Friedrich, *Répertoire métrique de la poésie lyrique française des origines à 1350*, Munchen, W. Fink Verlag, 1972.
- Naetebus, Gotthold, *Die Nicht-Lyrischen Strophenformen Des Altfranzösischen. Ein Verzeichnis Zusammengestellt Und Erläutert*, Leipzig, S. Hirzel, 1891.
- Pagnotta, Linda, *Repertorio metrico della ballata italiana*, Milano; Napoli, Ricciardi, 1995.
- Parramon i Blasco, Jordi, *Repertori mètric de la poesia catalana medieval*, Barcelone, Curial, Abadia de Montserrat, 1992 (Textos i estudis de cultura catalana, 27).
- Solimena, Adriana, *Repertorio metrico dei poeti siculo-toscani*, Centro di studi filologici e linguistici siciliani in Palermo, 2000.
- Solimena, Adriana, *Repertorio metrico dello Stil novo*, Roma, Presso la Societa, 1980.
- Tavani, Guiseppe, *Repertorio metrico della lingua galego-portoghese*, Roma, Edizioni dell'Ateneo, 1967.

Bibliography on Spanish metrical studies

- Baehr, Rudolf, *Manual de versificación española*, Madrid, Gredos, 1970.
- Balbín, Rafael de, *Sistema de rítmica castellana*, Madrid, Gredos, 1968.
- Beltrán, Vicenç, *Bibliografía sobre poesía medieval y cancioneros*, publicada en la Biblioteca Virtual "Joan Lluis Vives" *http://www.lluisvives.com/*, 2007.
- Bonnín Valls, Ignacio, *La versificación española. Manual crítico y práctico de métrica*, Barcelona, Ediciones Octaedro, 1996.

- Domínguez Caparrós, José, *Diccionario de métrica española*, Madrid, Paraninfo, 1985.
- _____, *Métrica y poética. Bases para la fundamentación de la métrica en la moderna teoría literaria*, Madrid, U.N.E.D., 1988a.
- _____, *Contribución a la bibliografía de los últimos treinta años sobre métrica española*, Madrid, C.S.I.C., 1988b.
- _____, *Métrica española*, Madrid, Síntesis, 1993.
- _____, *Métrica comparada: española, catalana y vasca. Guía didáctica*. Madrid, U.N.E.D., 1994.
- _____, *Estudios de métrica*, Madrid, U.N.E.D., 1999.
- _____, *Análisis métrico y comentario estilístico de textos literarios*. Madrid, Universidad Nacional de Educación a Distancia, 2002.
- Duffell, Martin, *Syllable and Accent: Studies on Medieval Hispanic Metrics*, Londres, Queen Mary and Westfield College, 2007.
- García Calvo, Agustín, *Tratado de Rítmica y Prosodia y de Métrica y Versificación*, Zamora, Lucina, 2006.
- Gómez Redondo, Fernando, *Artes Poéticas Medievales*, Madrid, Laberinto, 2001.
- González-Blanco García, Elena, *La cuaderna vía española en su marco panrománico*, Madrid, FUE, 2010.
- _____ y Seláf, Levente, "*Megarep*: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires", *Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world*, eds. L. Soriano - M. Coderch - H. Rovira - G. Sabaté - X. Espluga. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang, 2013.
- Herrero, José Luis, *Métrica española. Teoría y práctica*, Madrid, Ediciones del Orto, 1995.
- Mario, Luis, *Ciencia y arte del verso castellano*, Miami, Universal, 1991.
- Micó, José María, *Bibliografía para una historia de las formas poéticas en España*, ed. Digital, Biblioteca Virtual Miguel de Cervantes, Alicante, 2009 [*www.cervantesvirtual.com*]
- Navarro Tomás, Tomás, *Métrica española. Reseña histórica y descriptiva*, Syracuse, Syracuse University Press, 1956.

- _____, *Arte del verso*, México, Compañía General de Ediciones, 1959.
- _____, *Repertorio de Estrofas Españolas*, New York, Las Americas Publishing Company, 1968.
- Paraíso, Isabel, *La métrica española en su contexto románico*, Madrid, Arco Libros, 2000.
- Quilis, Antonio, *Métrica española*, Madrid, Ediciones Alcalá, 1969.
- Seláf, Levente, *Chanter plus haut. La chanson religieuse en langues vernaculaires. Essai de contextualisation*, Champion, 2009.
- Torre, Esteban, *El ritmo del verso: estudios sobre el cómputo silábico y distribución acentual a la luz de la métrica comparada*, Murcia, Universidad de Murcia, Servicio de Publicaciones, 1999.
- _____, *Métrica española comparada*, Sevilla, Servicio de Publicaciones de la Universidad de Sevilla, 2000.
- Utrera Torremocha, *Historia y teoría del verso libre*, Ed. Padilla libros, 2001.
- Valero Merino, Elena, Moíno Sánchez, Pablo y Jauralde Pou, Pablo, *Manual de métrica española*, Madrid, Castalia, 2005.

TEI-conform XML Annotation of a Digital Dictionary of Surnames in Germany

Horn, Franziska; Denzer, Sandra

In this paper we focus on XML markup for the *Digital Dictionary of Surnames in Germany* (*Digitales Familiennamenwörterbuch Deutschlands*, DFD). The dictionary aims to explain the etymology, and the meaning of surnames respectively, occurring in Germany. Possibilities and constraints which are discussed can be stated by using the TEI module

"Dictionaries" for editing a specialized dictionary such as the DFD. This topic includes situating the new project within the landscape of electronic dictionaries.

Our evaluation of the appropriateness of the proposed guidelines is seen as a contribution to the efforts of the TEI: The consortium regards their specifications as dynamic and ongoing development. The efforts in terms of lexical resources starting with the digitization of printed dictionaries are documented and discussed in various publications (e.g. Ide/Véronis/Warwick-Armstrong/Calzolari 1992; Ide/Le Maitre/Véronis 1994; Ide/Kilgarriff/Romary 2000). The module “Dictionaries” contains widely accepted proposals for digitizing printed dictionaries but projects which are born digital are progressively becoming more common nowadays (Budin/Majewski/Mörth 2012). For a more fine-grained encoding of these resources certain proposals for customization of the module "Dictionaries" can be found (e.g. Budin/Majewski/Mörth 2012). This paper aims to focus on the usefulness of the guidelines for a dynamic and specialized online dictionary without customized TEI extensions. Yet, our investigation points out possible extensions which may increase the acceptance and application of the TEI in other, similar projects.

At first, we want to introduce the *Digital Dictionary of Surnames in Germany* (2012-2036) as a new and ongoing collaboration between the Academy of Science and Literature in Mainz and Technische Universität Darmstadt. Work on DFD started in 2012. The project is based on data of the German telecommunications company Deutsche Telekom AG and preliminary studies of the *German Surname Atlas* (*Deutscher Familiennamenatlas*, DFA). It is planned to integrate the dictionary in an online portal of onomastics named *namenforschung.net* which can be seen as a gateway to various projects and information related to the field of name studies.

The intention of the DFD is to record the entire inventory of surnames occurring in Germany including foreign ones. Therefore, the entries consist of several features, for instance frequency, meaning and etymology, historical examples, variants and the distribution of the surnames. The short introduction includes a brief classification of the DFD into a typology of dictionaries (Kühn 1989; Hausmann 1989). Then,

we focus on data annotation in terms of the DFD according to the TEI Guidelines as the consortium forms a de facto standard for the encoding of electronic texts (Jannidis 2009). Following the proposals means providing possibilities for data exchange and further exploration (Ide/Sperberg-McQueen 1995). Both aspects are particularly important considering the long duration of the project. The encoding scheme of the DFD is mainly based on the TEI module "Dictionaries". Furthermore, components of the modules "Core" as well as "Names, Dates, People, and Places" are used. The main reason for considering the latter module is the close connection of surnames to geographical features, for example settlements or rivers. TEI extensions for customizing existing tags and annotation hierarchies according specific needs are set aside to provide a higher level of data interchangeability, for instance with other TEI and XML-based onomastic projects such as the *Digitales Ortsnamenbuch Online* (DONBO), a digital dictionary of place names (Buchner/Winner 2011).

To evaluate the appropriateness of the TEI Guidelines regarding to our project we compare them to the needs of annotating microstructures of the DFD entries. The intention of the TEI is to offer exact as well as flexible annotation schemes (Ide/Sperberg-McQueen 1995). Therefore, relevant criteria for the evaluation are the completeness of the tagset and the flexibility in arranging elements and attributes. Furthermore, the analysis discusses the comprehensibility of possible annotations in terms of descriptive and direct denotations.

In general, the TEI Guidelines – the tagset and the arrangement of its elements – can be used to represent the structure of the entries as well as the features of the DFD adequately. The applicability is, however, influenced by several aspects we want to discuss in greater detail.

At first, the aspect of completeness of the tagset is discussed. It would be useful to have elements within the module "Dictionaries" available to encode the frequency and the geographical distribution. The frequency of a surname is interesting for dictionary users, especially the name bearer. Other than for the DFD, options to encode frequencies seem to be important considering other lexical resources such as explicit frequency dictionaries or the frequency information in learner's dictionaries, for instance. Elements to annotate the geographical distribution are needed,

because the distribution in and outside of Germany serves as means to support, respectively verify, the given sense-related information (Schmuck/Dräger 2008; Nübling/Kunze 2006). These tags seem to be of further interest for parallel developments of national surname dictionaries, for example in Austria (*FamOs*) as well as for other types of dictionaries, for instance, variety dictionaries.

In our encoding scheme, the missing tags are replaced by more indirect combinations of tags und attributes, for example <usg type="token"> to encode the frequency or <usg type="german_distribution"> to annotate the distribution.

Furthermore, it would be helpful to have more possibilities to specify a sense. According to the presentation of surnames in the DFD, a sense is linked with a category, which can be understood as a type of motivation for the given name. An example is the category *occupation* belonging to the surname *Bäcker* ('baker'). For our purposes it is adverse that the attribute @type is not allowed within the element <sense>. We are using the less concise attribute @value as an alternative.

A further example for missing options of explicit markup relates to the sense part. In the DFD senses are ordered according to their certainty. We are using the attribute @expand with the values "primary", "uncommon", "uncertain" and "obsolete" to differentiate. However, the definition provided by the TEI Guidelines entails giving an expanded form of information (TEI Consortium P5 2012). The slightly different usage in the DFD annotation scheme is based on the lack of suitable alternatives and the denotative meaning of the expression to expand. Furthermore, it would be helpful to have elements within the module "Names, Dates, People, and Places" which encode not only settlements, place names and geographical names in general but more precise features as hydronyms or agronyms, additionally. Currently, these features are tagged as follows in our articles: <geogName type="hydronym"/>. Another aspect is the indefinite usage of one element in several contexts. An example is the tag <surname> which can be used to encode the surname in general as well as to annotate the explicit last name of a certain author of a cited publication.

The appropriateness of the module "Dictionaries" for encoding the DFD is diminished by restrictions concerning the arrangement of elements.

The element <bibl> for annotating bibliographic references is not allowed on the entry or sense level. Within the project *Wörterbuchnetz*, the restriction in terms of the sense-element is overridden by embedding the element <bibl> within the element <title> or <cit> (Hildenbrandt 2011). The encoding scheme of the DFD uses the element <cit> as TEI-conform parent-element. For example: `<cit> <bibl> <author> <surname>Gottschald</surname> </author> <date when="2006"/> <biblScope type="pp">5</ biblScope> </bibl> </cit>`

The risk of these flexible solutions is that similar projects might handle similar situations by choosing different TEI-conform markup strategies or customizations by TEI extensions which limits the possibilities for interchange.

As a result, we find that some aspects are not as adequately considered within the TEI modules "Dictionaries" and "Names, Dates, People, and Places" as it would be useful to realize the intended function of a new dictionary of surnames in Germany. An extension of the tagset might include elements for the frequency and the distribution. A further proposal refers to the element <bibl>, which should be allowed in more contexts. The pursuit of the TEI Guidelines, which is to provide an expressive and explicit tagset, is not fulfilled completely in terms of the DFD: The indirect denotations and the vast usage of attributes affect the readability for human lexicographers working on the XML adversely. These are among the reasons for the development of a working environment using the author view of the xml editor Oxygen instead of the source view.

Our explanations might give impetus for slight extensions of the TEI to develop a more comprehensive, comprehensible and flexible annotation scheme for general dictionaries as well as a more adequate annotation scheme for specialized dictionaries. An appropriate and profound encoding can be seen as the basis for an abundance of application scenarios of the DFD.

Bibliography

- Austrian Academy of Sciences (ed.) (n.d.) Familiennamen Österreichs (FamOs). http://hw.oeaw.ac.at/famos (accessed June 30, 2013).

- Buchner, S./Winner, M. (2011). Digitales Ortsnamenbuch (DONBO). Neue Perspektiven der Namenforschung. In Ziegler, A./Windberger-Heidenkummer, E. (eds.): Methoden der Namenforschung. Methodologie, Methodik und Praxis. Berlin: Akademie Verlag, pp. 183-198.
- Budin, G./Majewski, S./Mörth, K. (2012). Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. In Journal of the Text Encoding Initiative. 3. November 2012, Online since 15 October 2012. URL: http://jtei.revues.org/522; DOI: 10.4000/jtei.522. (accessed June 30, 2013).
- Hausmann, F. J. (1989). Wörterbuchtypologie. In Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher: Ein internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter, pp. 968-980.
- Hildenbrandt, V. (2011). TEI-basierte Modellierung von Retrodigitalisaten (am Beispiel des Trierer Wörterbuchnetzes). In Klosa, A./Müller-Spitzer, C. (eds.): Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie". Mannheim: Institut für Deutsche Sprache, pp. 21-35.
- Ide, N./Kilgarriff, A./Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In Proceedings of Euralex 2000. Stuttgart, 113-126.
- Ide, N./Le Maitre, J./Véronis, J. (1994). Outline of a Model of for Lexical Databases. In Zampolli, A./Calzolari, N./Palmer, M. (eds.): Current Issues in Computational Linguistics. Pisa: Giardini Editori, pp. 283-320.
- Ide, N./Sperberg-McQueen, M. (1995). The TEI. History, Goals, and Future. In Computers and the Humanities 29, 5-15.
- Ide, N./Véronis, J./Warwick-Armstrong, S./Calzolari, N. (1992). Principles for encoding machine readable dictionaries. In Tommola, H./Varantola, K./Salmi-Tolonen, T./Schopp, Y. (eds.): EURALEX '92. Pproceedings I- II. Papers submitted to the 5th

EURALEX International Congress on Lexicography in Tampere, Finland. Tampere: Tampereen Yliopisto, pp. 239-246.

- Jannidis, F. (2009). TEI in a Crystal Ball. In Literary and Linguistic Computing. 24(3), 253-265.
- Kühn, P. (1989). Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher: Ein internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter, pp. 111-127.
- Nübling, D./Kunze, K. (2006). New Perspectives on Müller, Meyer, Schmidt: Computer-based Surname Geography and the German Surname Atlas Project. In Studia Anthroponymica Scandinavica. Tidskrift för nordisk personnamnsforskning 24, 53-85.
- Schmuck, M./Dräger, K. (2008). The German Surname Atlas Project. Computer-Based Surname Geography. In Proceedings of the 23rd International Congress of Onomastic Sciences. Toronto, 319-336.
- TEI Consortium (eds.). Guidelines for Electronic Text Encoding and Interchange. 17th January 2013. http://www.tei-c.org/P5/ (accessed June 30, 2013).
- Trier Center for Digital Humanities (ed.) (n.d.) Wörterbuchnetz. http://woerterbuchnetz.de/ (accessed June 30, 2013).

From Paper Browser to Digital Scientific Edition of Ancient Written Sources

Lamé, Marion; Kossman, Perrine

To this day, digital epigraphy has developed following two paths. The first one corresponds to the reproduction of the way information

is structured in a corpus published on paper, with an additional browsing, search, and data extraction option. For instance, databases like Inscriptions of Aphrodisias On Line (http://insaph.kcl.ac.uk/iaph2007/index.html), Clauss-Slaby (http://www.manfredclauss.de/fr/), EDH (http://edh-www.adw.uni-heidelberg.de/home), or ultimately Phi7 (http://epigraphy.packhum.org/inscriptions/), in spite of the fact that they are efficiently searchable, are structured simply according to the traditional elements of a paper publication, that is to say lemma, diplomatic transcription, critical edition, typographical code, translation, apparatus criticus, historical commentary. The steps involved in their consultation somehow reproduce the ones taken in the consultation of a paper edition in a library, except that the search is quicker and more powerful: one click of the mouse opens a related map or dictionary entry. This is why we are tempted to call such information systems "paper browsers". But digital scientific editions have more to offer, and some projects have already explored another path, resulting in an attempt to go beyond the possibilities of a paper publication. The best example of this trend is the well known website Res Gestae Divi Augusti Fotogrammetria http://resgestae.units.it/index.jsp, which allows to browse a digital version of highly reliable interactive photogrammetric photographs and squeezes of the huge inscription, that are reasonably impossible to print on paper. Another instance would be the ChiSel System (http://chisel.hypotheses.org/tag/presentation?lang=es_ES), which generates 3d representations of written objects.

Such achievements lead the way to a new kind of information systems, not based on the digitization of the epigraphic knowledge as it is published on paper anymore. On the contrary, a new conceptual model is required: a model that would revert to what an inscription really is, and thus would be able to fully exploit the abilities of the digital environment to express its multidimensional aspects. Ideally, it should be collectively defined.

Following that idea, in this poster we would like to focus more specifically on the textual aspects of the digital representation of inscriptions, expressed in "paper browsers" via the subset EpiDoc TEI (rethinking diplomatic and critical organisation in levels 5 and 6 of Lamé & Valchera 2012). Using a methodological and at the same time experimental

approach, as McCARTHY 2005, and before him BORILLO, 1984, encourages, we would like to demonstrate that EpiDoc TEI, whereas it has developed along the "paper browsers" experience, offers more possibilities and can perfectly fulfill the needs of a digital edition as briefly defined previously, if it takes into account the real epigraphic object in all its dimensions (writing, context...).
Thanks to three case studies, we hope to demonstrate its current capacities and what its best use could be. First, we will try and construct the standoff position of a bilingual inscription from Samos (Demotic and Greek texts); then the standoff position of the partly preserved dedication on a statue base also from Samos; and finally the standoff position of two Roman inscriptions, CIL, 11, 6664, CIL, 11, 1421, particularly interesting for the entangled abbreviations, stuck words, mistakes and ligatures. Hopefully, those analyses will help determine how TEI could be optimally used.
We hope that this poster will create the opportunity of a dynamic and fruitful discussion with the TEI community.

Bibliography

Digital humanities bibliography

- BORILLO, M. 1984 *Informatique pour les sciences de l'homme* Bruxelles Mardaga
- CIOTTI, F. 2005 'La codifica del testo XML e la Text Encoding Initiative' *Il manuale TEI Lite: Introduzione alla codifica elettronica dei testi letterari* Milano Sylvestre Bonnard 9-42
- GENET, J.-P. 1994 'Source, métasource, texte, histoire' *Storia & multimedia: Atti Del Settimo Congresso Internazionale Association for History & Computing* Bologna Grafis 3–17
- FUSI, D. 2007 'Edizione epigrafica digitale di testi greci e latini: dal testo marcato alla banca dati' *Digital Philology and Medieval Texts*Ospedaletto (Pisa) Pacini pp. 121–163
- FUSI, D. 2011 *Informatica per le scienze umane Vol. 1 - Elementi* Roma Nuova Cultura 1
- FUSI, D.2011 *Informatica per le scienze umane Vol. 2 - Modelli* Roma Nuova Cultura 2

- GOLD, M. 2012 *Debates in the digital humanities* Minneapolis University of Minnesota Press
- GREENGRASS, M., & LORNA, H. 2008 *The virtual representation of the past* Farnham Ashgate
- LAMÉ, M., VALCHERA, V., & BOSCHETTI, F. 2012 'Epigrafia digitale#: paradigmi di rappresentazione per il trattamento digitale delle epigrafi' Epigraphica 386–392
- LUNEFELD, P., BURDICK, A., DRUCKER, J., PRESNER, T., & SCHNAPP, J. 2012 *Digital_Humanities* Boston MIT Press
- McCARTHY, W. 2005 *Humanities Computing* New York Palgrave Macmillan
- NEROZZI-BELLMAN, P. 1997 *Internet e le muse: la rivoluzione digitale nella cultura umanistica* Milano Associazione Culturale Mimesis
- ORLANDI, T. 1985 'Problemi di codifica e trattamento informatico in campo filologico' Lessicografia, Filologia e Critica Firenze Leo S. Olschki 42 69-81
- PERILLI, & FIORMONTE, D. 2011 *La macchina del tempo. Studi di informatica umanistica in onore di Tito Orlandi* Firenze Le Lettere
- PIERAZZO, E. 2005 *La codifica dei testi* Roma Carocci
- RONCAGLIA, G. 1997 'Alcune riflessioni su edizioni critiche, edizioni elettroniche, edizioni in rete' *Internet e Le muse: La Rivoluzione Digitale Nella Culture Umanistica* Milano Associazione Culturale Mimesis251–276
- ROUECHÉ, C. 2009 *Digitizing Inscribed Texts». In: Text Editing, Print and the Digital World* Farnham Ashgate 159–168
- SMITH, N. 2012 'Les-humanités-dont-on-ne-doit-pas-prononcer-le-nom' Translated byM. Lamé *Read / Write Book 2* P. Mounier Open Edition press 87–88 http://vitruviandesign.blogspot.it/2012/01/humanities-that-must-not-be-named.html
- SOLER, F.From 2012 *Carnet de recherche Chisel* http://chisel.hypotheses.org [37]

[37] Carnet de recherche sur la plateforme *Hypotheses.org*

- SUSINI, G. 1982 *Epigrafia romana* Roma Jouvence
- TORRES, J.C., SOLER, F. 2012 'An Information System to Analize Cultural Heritage InformationPaper accepted Euromed Conference 2012 '

Digital humanities bibliography

- BELLET, M.-É. & al. 2003 *De la restitution en archéologie. Actes du Colloque de Béziers organisé par le Centre des Monuments nationaux* Paris Éditions du Patrimoine http://editions.monuments-nationaux.fr/fr/le-catalogue/bdd/livre/662
- ÉTIENNE, R. 1970 *Le siècle d'Auguste* Paris Armand Colin
- GHINATTI, F. 1999 *Alfabeti greci. Torino: Paravia scriptorium*
- JACQUES, F. 1990 *Les Cités de l'Occident romain* Paris Belles Lettres
- KRUMMREY, H., & PANCIERA, S. 1980 'Criteri di edizione e segni diacritici' Tituli 2: Miscellanea Roma Edizioni di storia e letteratura 2 205–215
- PANCIERA, S. 2012 *What Is an Inscription? Problems of Definition and Identity of an Historical Source* Translated byJ. BODEL *Zeitschrift für Papyrologie und Epigrafik* 183 1-10
- ROW, G. 2002 *Princes and Political Culture* Ann Arbor University of Michigan Press

Sources Edition

- ARIAS, P.E., CRISTANI, E., GABA, E. 1977 *Camposanto monumentale di Pisa* Pisa Pacini
- HALLOF, KI. 2000 *n° 348* Inscriptiones Graecae XII 6, I
- HALLOF, KI. 2003 *n° 589* Inscriptiones Graecae XII 6, II
- LUPI, C. 1979 *I decreti della colonia pisana ridotti a miglior lezione* Pisa F. Mariotti e CC.
- MAROTTA D'AGATA, R. 1980 *Decreta Pisana (CIL, XI, 1420-21)ed. critica, trad. e commento* Pisa Ed. Marlin
- SEGENNI, S. 2011 *I Decreta Pisana : autonomia cittadina e ideologia imperiale nella colonia Opsequens Iulia Pisana* Bari Edipuglia

A Challenge to Dissemination of TEI among a Language and Area: A Case Study in Japan

Nagasaki, Kiyonori; Muller, Charles; Shimoda, Masahiro

This presentation describes a challenge to the dissemination of TEI in Japan, a country where most of the people have spoken and written in a single language for more than a millennium. There are at present, very few examples of attempts at adopting TEI for Japanese cultural resources. However, there has been a rich textual tradition, and many textual resources are preserved going back as far as the 8th century. A vastly greater amount of materials remain from the 17th century, due to the spread of technologies of woodblock printing. Humanities researchers have addressed the digitization of humanities resources since the 1950's.

In the early stages, Japanese linguists began attempts at digitizing language resources in order to statistically analyze Japanese and Western materials, publishing a journal through the establishment in 1957 of a society named "Mathematical Linguistic Society of Japan" (Keiryo Kokugo Gakkai, ######)[38]. In addition, several progressive researchers working at the National Institute for Japanese Language and Linguistics, National Institute for Japanese Literature, and several universities commenced the digitization of their Japanese materials using large-scale computer systems. The National Museum of Ethnology also played an important role in this endeavor.

Following upon these early attempts, several communities were established at the end of the 1980's due to the impetus of the proliferation of the IBM PC. One was formed as the Special Interest Group of Computers and the Humanities[39], that is, SIG-CH, under the auspices of the Information Processing Society in Japan, the largest computer science society in Japan. The others were the Japan Society of Information and Knowledge[40] and the Japan Art Documentation Society[41]. After

[38] http://www.math-ling.org/e-index.html

[39] http://www.jinmoncom.jp/

[40] http://www.jsik.jp/?index-e

that, many academic communities were established based on the new possibilities opened up by the Internet. It is especially noteworthy that societies of digital scholarship of archaeology, English corpora, and Asian literature were formed in the 1990's. Moreover, several academic communities have been formed even in the 21st century, including JADH (Japanese Association for Digital Humanities)[42] which has become a constituent organization of ADHO.

Under these circumstances, over a thousand presentations regarding digitization of the humanities have been made since the 1950's. Around 800 presentations have been done in quarterly workshops of SIG-CH from 1989 to 2012, including various types of digital scholarship in the humanities such as textual analysis, text database, image database, and so on (Figure.1), targeting various fields in the humanities (Figure 2)

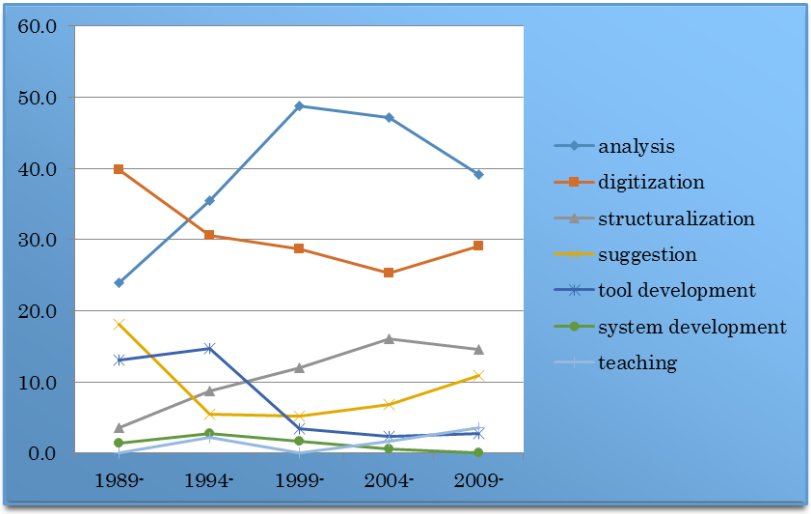


Figure 1. Types of digital scholarship in the presentations of SIG-CH

[41] http://www.jads.org/eng/index.html

[42] http://www.jadh.org/

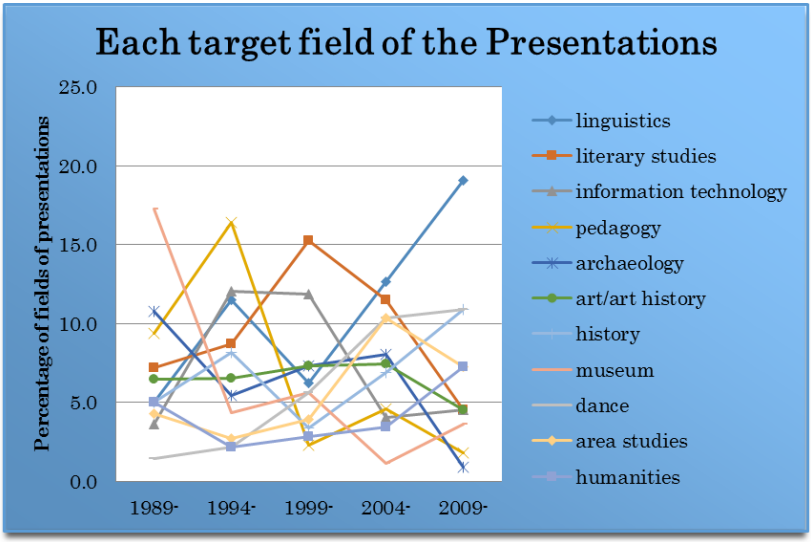


Figure 2. Top 11 target fields of the presentations

However, the TEI has not fared that well up to now in Japanese academic communities--probably due to several reasons, including the issues of character encoding and language barriers. Actually, differences in character encoding prevented sharing of a broad range of digital content and applications beyond TEI. Many of the applications that were developed for western languages could not be used under Japanese computer environments before the promulgation of Unicode. This means that it was difficult for Japanese humanities researchers to realize the significance and potential of TEI at that time. Moreover, it was also difficult to participate in the discussion of TEI due to difference of language and the large distance from the center of TEI. Therefore, in spite of efforts of few researches, Japanese researchers had rarely participated in the activities of TEI until recently. Instead, they had addressed their textual resources using their own individual approaches.

Recently, the pervasive implementation of Unicode and spread of the Internet widen the possibilities of TEI even in Japan. In 2006, a TEI meeting[43] hosted by Christian Wittern at Kyoto University gathered

[43] http://coe21.zinbun.kyoto-u.ac.jp/tei-day/

various researchers, newly awakening scholars to the potential of TEI. After that, a series of DH workshops including TEI tutorials in 2009 at Tokyo and Osaka began to be held by a DH community which led to the formation of an association called the Japanese Association for Digital Humanities later. In this new period, even in Japan, researchers of the humanities could experience the potential and possibilities of TEI by hands-on usage of several strong tools based on UTF-8 which were developed by TEI communities such as oXygen, Versioning Machine, Roma, and so on. These efforts were strongly supported by TEI specialists such as Espen Ore, John Lavagnino, Susan Schreibman, and Elena Pierazzo.[44] Several DH courses in Japanese universities have recently included tutorials on TEI.

Also, a project of Japanese translation of the TEI guidelines has been initiated by several young researchers led by Kazushi Ohya. Thus, the environment for TEI has been gradually forming in Japan. Actually, several DH projects are trying to use TEI for their digital resources. Their results will be shown in the near future.

During the discussion of adopting TEI on Japanese textual materials, several problems have been recognized. For example, Japanese texts often contain intralinear text that indicates phonetic representation called "ruby," which was already adopted in HTML5[45] and ePub 3.0.[46] It is not simply a phonetic standard, but its system can depends on the idiosyncratic phonetic representations of a certain author, editor, or publisher. Rather, it represents a phonetic rendering in specific situations. Probably type attributes can be applied in this case, but a guideline should be prepared for such usage. Otherwise, a module may need to be created specifically for handling Japanese materials. This kind of effort could be useful for dissemination of TEI in other countries and areas. Moreover, as already discussed in several places--such as DH2012, some linguists would prefer to avoid using not only TEI but also general tags (even in Japan) so that

[44] Most of the information of the workshops are put on the JADH Web site. (http://www.jadh.org). This series yielded a Web page "A Simple Guide to TEI and oXygen (in Japanese)" which are referred in various related workshops in Japan. (http://www.dhii.jp/nagasaki/blog/node/12)

[45] http://www.w3.org/TR/html5/text-level-semantics.html#the-ruby-element

[46] http://www.idpf.org/epub/30/spec/epub30-contentdocs.html

they can mine texts freely. We should discuss this matter carefully and constructively.

Finally, stand-off markup seems to be suitable for most Japanese resources, but meticulous application has not been carried out so far. It should be solved as soon as possible.

Bibliography

- Kiyonori Nagasaki, How Digital Technologies Have Been Used: Through the History of the SIG Computers and the Humanities, "IPSJ technical report", 2013-CH-98(7), pp. 1-6. (in Japanese)
- Kiyonori Nagasaki, "A Simple Guide to TEI and oXygen", [http://www.dhii.jp/nagasaki/blog/node/12] (in Japanese)

Dramawebben, linking the performing arts and the scholarly communities

Olsson, Leif-Jöran; Forsbom, Eva; Lagercrantz, Marika; Lindgren, Ulrika

Background

Dramawebben (The Swedish Drama Web) has served as a free digital resource since 2006. A largely unexplored empirical material of Swedish drama free from copyright has been made accessible through a website [1]. The website has been used by scholars and students, theatre practitioners and a general public.

In most cases the first printing of the play was the version first encountered by a theatrical audience. First printings are also the most difficult to access and thus the most exclusive editions and therefore the most important to make accessible. Plays are generally published in two formats: a facsimile and a text version (made from optical character recognition of

the facsimile images). They are also accompanied by descriptive catalogue entries, where a reading of each individual play, and of its reception in the press at its first performance, are summarised in informative meta texts. This publication principle is an important preparatory foundation for scholarship, where the facsimile functions as a complement to the encoded text version. Being able to switch between facsimile and encoded text versions is sometimes important, e.g. when the text is in Gothic type or the text only exists in the form of a handwritten manuscript.
Each stage of the work of collecting, processing and publishing the material has been designed in such a way as to lay a preparatory foundation for scholarship that will hold for a development of Dramawebben into an exemplary national infrastructure for digital research in the humanities. In an ongoing project 2012-2014, Dramawebben is further developing the website, making a foundation for pushing the e-Drama infrastructure into a long-term operation. The project includes a baselined corpus of TEI-drama annotated plays and development of exploration tools, and engaging a vibrant community. A key component is to educate students in TEI-encoding and let them be ambassadeurs spreading the word to target disciplines within the humanities, such as linguistics, literary and theatre history, studies in children's culture, practical and theoretical research in children's theatre, and arts tertiary institutions.

Collaboration and sustainability

Since its start in 2006, Dramawebben has initiated collaboration with a number of other infrastructures for the mutual benefit of all parties involved. Such cooperation is also ensuring long-term sustainability, and that the research material will be available as a national resource.
In cooperation with Språkbanken (The Swedish Language Bank), tools for linguistic annotations, search functions and display formats for linguistic investigations will be available for Dramawebben [2][3]. Språkbanken, on the other hand, can include drama, an otherwise missing text type, in their language corpora. Språkbanken is also involved in Litteraturbanken (The Swedish Literature Bank), another digitisation infrastructure. Litteraturbanken has provided Dramawebben with advice and technical support on standards for digitisation, publication and process support, and Litteraturbanken uses facsimiles made by Dramawebben.

Making the material accessible has always been a high priority for Dramawebben. Therefore, Dramawebben is included in the search engines Libris and K-samsök of Kungliga biblioteket (The National Library of Sweden) and Riksantikvarieämbetet (The Swedish Central Board of National Antiquities), respectively. There are links from the catalogue entries to the library databases, which refer users to the original material in each respective archive, while, for example, Libris links to Dramawebben's entries in order to refer users to meta data and full text publications. It is planned that Dramawebben will be included in Libris as its first digital archive.

Dramawebben has also been conducting a very fruitful collaboration in the field of digitisation with the National Library of Sweden, and the archives of the Royal Opera, Royal Dramatic Theatre, and Statens musikverk (Music Development and Heritage Sweden). Through supplementary grants from the Bank of Sweden Tercentenary Foundation, 15,000 pages of printed drama from the national library's collections have been photographed by the library's own digitisation department. In the ongoing project, digitisation of handwritten material in the theatrical archives are being made, to the benefit of all parties concerned.

TEI-drama encoding

In order to facilitate more advanced exploration within and across dramas, we are in the process of TEI-encoding a subset of the plays. By adhering to TEI text encoding principles, we make a commitment to sustainability, but can also benefit from being part of a larger community. Preparation for text encoding started in the spring of 2012. All plays on Dramawebben printed 1880-1900 were selected. It included 89 plays in all genres, children's plays, drama and comedy, plays by female as well as by male dramatists.

Baseline encoding

Common for all plays is a baseline encoding taken from the drama module of TEI, and minimal support for facsimile encoding, connecting the TEI-encoded text to the facsimile. The baseline encoding covers the basic structure of the drama text. On top of that, it is possible to add semantic annotation, which goes beyond the text itself, referring to the action below, behind or beyond the actual words.

Semantic encoding

To tempt scholars in humanities with at least one theme for semantic encoding, we have started with one – textile handicraft, which was a recurrent feature of the plays by female playwrights of the 1880's. The needle working woman was a strong and yet ambivalent sign from the period. August Strindberg let one of his heroines deny the crochet she was constantly working on: It is nothing, it is just my needlework[4]. To his female colleague Alfhild Agrell the handicraft had a subversive power. One of her heroines silently embroidered her way to financial independence and freedom from an unbearable marriage[5]. Strindberg's heroine denies her needlework but still performs it in full limelight. Needlework it is a potent stage action or, a playable sign.

So how did we go about encoding this manifold sign? We soon realized it was not always fully designated in the stage directions. Although the props and starting point of the action was given - She picks her knitting – the point where the action ceases might not always be mentioned. The question when she puts down the knitting can be related to why she quits. Encoding handicraft thus opens to the exploratory reading of the drama text that is the basis for every stage action. And it is in this very interpretative process that scholars will meet theatre practitioners.

The textile handicraft is not only embodied in the actual stage action. It will also be present in the lines where the speakers elaborate their knowledge and attitude about it. The props as well as the handicraft are also frequently used as metaphors for life, death and fate as well as for daily matters.

Dissemination

Our task is not only to do the text encoding, but also to implement and spread TEI as a new research tool in the Swedish communities of humanities and of artistic production, by employing students as ambassadeurs. We apply an adapted version of the bottom-up process practiced by the Women Writers Project at Browns[6], meaning that digital humanities must come from the grass roots – the students. The TEI-encoding is therefore performed by five students in literature and theatre science, simultaneously functioning as ambassadeurs.

They have assimilated TEI and the basic encoding quickly. During the first five months of approximately 100 hours work they have also increased substantially in speed and accuracy. That has been a process of finding their own way of balancing transcribing, encoding and proof reading. The students have been encouraged to not only perform the basic encoding but also find their own themes for semantic encoding.

Three workshops will be held during 2013-2014, where the students and invited scholars will present their explorations into the potentials and adventures of digital humanities, given their respective use cases. Main target groups are scholars, theatre practitioners and librarians, who are not familiar with the possibilities of TEI-encoding.

Acknowledgements

The authors gratefully acknowledge financial support from the Swedish Research Council (VR Dnr: 2011-6202).

Bibliography

- Dramawebben <http://www.dramawebben.se>
- Korp, Språkbanken, University of Gothenburg, <http://spraakbanken.gu.se/korp/>.
- Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012. The open lexical infrastructure of Språkbanken. Proceedings of LREC 2012, Istanbul: ELRA. 3598-3602 <http://spraakbanken.gu.se/karp/>.
- August Strindberg, To Damascus, 1898.
- Alfhild Agrell, Saved, 1883.
- Women Writers Project, Brown University <http://www.wwp.brown.edu/>.

The Karnak Cachette Texts On-Line: the Encoding of Transliterated Hieroglyphic Inscriptions

Razanajao, Vincent; Morlock, Emmanuelle; Coulon, Laurent

Between 1903 and 1907, G. Legrain discovered around 800 stone statues, stelae and other objects in a large pit (the so-called "Cachette") inside the temple of Amun at Karnak, in which they were piously buried by the Egyptian priests, probably during the 1st century B.C. They include a number of royal effigies of all periods but most of the statues primarily belong to the priests who officiated at Karnak from the New Kingdom to the end of the Ptolemaic Period.

The Karnak Cachette Database is an on-line inventory of the Cachette and a tool to search this rich corpus. The first version was launched in 2009; it provides, insofar as possible, a general description of each object (with dimensions, materials, dating), a label, the date of discovery, different inventory numbers, and a bibliography. Version 2 was put online in 2012: it includes an extensive access to the photographic documentation (more than 8,000 photographs are now available); this database has been regularly updated thereafter.

Building on this well-defined corpus, the project aims now at developing the tools to encode, search and publish electronically the hieroglyphic texts inscribed on these objects, which provide anthroponomical, toponymical and prosopographical data and are therefore of historical and documentary significance. The encoding is developed according to the recommendations of the Text Encoding Initiative in combination with relevant "best practices" in the field of Digital Humanities applied to Epigraphy (Elliott et alii 2007; Cayless et alii, 2009). In this sense, even though the project takes into account many of the EpiDoc schema rules, it is only partially compliant with this TEI customization because of the specificities of both the project and Ancient Egyptian Epigraphy (compare with Lamé 2009), and also because there is a necessity to fall within the scope of other Egyptological projects dealing with textual corpora (Winand, Polis, Rosmorduc in press).

Xefee, a tool to encode transliterated hieroglyphic inscriptions

It is well known that XML is far from being a human friendly way to encode texts. Several XML editors are already available; some of them are highly customizable and can be used by very specific project, providing the users are proprely trained and some implementation time and effort is spent. However, due to the specific features of the texts from the Karnak Cachette – for instance in terms of prosopography –, and the general philosophy of the project – edit and analyse texts that require full Egyptological proficiencies –, it has been decided to create a specific XML Editor that would make easier the text input, its marking up as well as the generation of the XML/TEI files.

Xefee – XML Editor for Egyptian Epigraphy – is a desktop Java application developed on Netbeans. It mainly consists of a general user interface (GUI) which provides all the necessary tools for managing and encoding the ancient Egyptian texts as well as the descriptive data pertaining to the Karnak Cachette project. These tools range from an import module that directly converts to XML the hieroglyphic text transcriptions written according to Egyptological standards, to more complex components intended to manage genealogical data.

The tab dealing with text encoding offers to the user a panel of buttons, combo-boxes and other controls that facilitate the marking up the texts with tags pertaining to epigraphy (<lb/>, <cb/>, <gap/>, <sic/>, <supplied/> elements), onomastic (<persName/> element and <rs/> elements with specific @type such as “deity”, “deityEpithet”, “toponym”) and prosopography (<rs/> elements with specific @type such as “person”, “title”, “filiationMark”). To add a tag, the user simply has to select in the top view pane the text to be marked up, and to press the appropriate button on the right-hand half of the tab. Since the XML marking up can be quite dense, mainly because the texts the project is dealing with often consist in compact sequences of personal names and titles, a preview pane in the bottom of the tab renders the encoded strings with different kinds of surrounding or highlighting patterns.

The Ancient Egyptian way to present genealogical filiations also required to build up peculiar tools to handle this very important aspect of the

text contents. A tab of the GUI is dedicated to the creation of person's identities, whilst another one intends to manage the family links and generate the <relationGrp/> element.

The current stage of the Karnak Cachette Project relies on the object and museum data described in the version 1 of the related database and on the photographic material added in its version 2. In order to fully use this already existing material, as well as to store the new data created throughout the encoding of the texts, Xefee leans on a MySQL database in which these different kinds of data are merged. Organised around a main "document" table, the data is spread over eighteen tables, among which four are dedicated to data from version 1, and one to the encoded texts.

In order to make full use of this material in a XML perspective, a sixth and last tab of the GUI is dedicated to the creation of the XML/TEI files. By pressing the upper-left button, the user asks Xefee to pick up in the MySQL database all the needed pieces of information and to place them between the appropriate XML tags. This generates all the sections of a XML file, from the headers with the publication and bibliographic statements to the div elements dealing with the encoded texts. The newly created XML file will be then poured into a native XML eXist database in order to constitute the electronic corpus itself.

Bibliography

- CACHETTE DE KARNAK: L. Coulon, E. Jambon, Base de données Cachette de Karnak /Karnak Cachette Database launched in November 2009; version 2 updated in January 2012. Karnak Cachette Database (http://www.ifao.egnet.net/bases/cachette).
- Cayless et alii 2009: H. Cayless, Charlotte Roueché, T. Elliott, G. Bodard, "Epigraphy in 2017", in Digital Humanities Quarterly 3.1 (2009). Available online.
- Elliott et alii 2007: T. Elliott, L. Anderson, Z. Au, G. Bodard, J. Bodel, H. Cayless, Ch. Crowther, J. Flanders, I. Marchesi, E. Mylonas and Ch. Roueché, EpiDoc: Guidelines for Structured Markup of Epigraphic Texts in TEI, release 5, 2007. Available online.

- Lamé 2008: M. Lamé, “Pour une codification historique des inscriptions”, Rivista Storica dell'Antichità 38, 2008 (2009), p. 213-225. Available online.
- Winand, Polis, Rosmorduc in press: J. Winand, St. Polis, S. Rosmorduc, “Ramses. An Annotated Corpus of Late Egyptian”, in P. Kousoulis (eds), Proceedings of the Xth International Association of Egyptologists Congress (Rhodes, Mai 2008), Leuven, Peeters, in press. Available online

Edition Visualisation Technology: a simple tool to visualize TEI-based digital editions

Rosselli Del Turco, Roberto; Masotti, Raffaele; Kenny, Julia; Leoni, Chiara; Pugliese, Jacopo

The TEI schemas and guidelines have made it possible for many scholars and researchers to encode texts of all kinds for (almost) all kinds of purposes: from simple publishing of documents in PDF form to sophisticated language analysis by means of computational linguistics tools. It is almost paradoxical, however, that this excellent standard is matched by an astounding diversity of publishing tools, which is particularly true when it comes to digital editions, in particular editions including images of manuscripts. This is in part due to the fact that, while there’s still an ongoing discussion about what exactly constitutes a digital edition, available publications have significantly raised users’ expectations: even a simple digital facsimile of a manuscript is usually accompanied by tools such as a magnifying lens or a zoom in/out tool, and if there is a diplomatic transcription (and/or a critical edition) we expect to have some form of image-text linking, hot-spots, a powerful search engine, and so on. The problem is that all of this comes at a cost, and the different needs of scholars, coupled with the constant search for an

effective price/result ratio and the locally available technical skills, have a led to a remarkable fragmentation: publishing solutions range from simple HTML pages produced using the TEI style sheets (or the TEI Boilerplate software) to very complex frameworks based on CMS and SQL search engines.

The optimal solution to the long standing visualization problem would be a simple, drop-in tool that would allow to create a digital edition by running one or more style sheets on the TEI document(s). The TEI Boilerplate software takes this approach exactly: you apply an XSLT style sheet to your already marked-up file(s), and you're presented with a web-ready document. Unfortunately, this project doesn't cover the case of an image-based digital edition I presented above, which is why I had to look elsewhere for my own research: the Digital Vercelli Book project aims at producing an online edition of this important manuscript, and has been examining several software tools for this purpose. In the end, we decided to build a software, named EVT (for Edition Visualization Technology), that would serve the project needs and possibly more: what started as an experiment has grown well beyond that, to the point of being almost usable as a general TEI publishing tool. EVT is based on the ideal work flow hinted above: you encode your edition, you drop the marked up files in the software directory, and voilà: after applying an XSLT style sheet, your edition is ready to be browsed. More in detail, EVT builder's transformation system divides an XML file holding the transcription of a manuscript into smaller portions each corresponding to individual pages of the manuscript, and for each of these portions of text it creates as many output files as requested by the file settings. Using XSLT modes to distinguish between the rules it is possible to achieve different transformations of a TEI element and to recall more XSLT stylesheets in order to manage the transformations. This allows to extract different texts for different edition levels (diplomatic, diplomatic-interpretative, critical) on the basis of the same XML file, and to insert them in the HTML site structure which is available as a separate XSLT module. If the TEI elements that are processed are placed in an HTML element with the class edition_level- TEI_ element's_name (e.g. for the element <abbr> in the transformation to the diplomatic edition: dipl-abbr) it is possible to

keep the semantic information contained in the markup and, if necessary, associate the element with that class of the CSS rules so as to specify the visualization and highlighting of the item. The edition level outputs and other aspects of the process can be configured editing the evt_builder-conf.xsl file.

At the present moment EVT can be used to create image-based editions with two possible edition levels: diplomatic and diplomatic-interpretative; this means that a transcription encoded using elements of the TEI transcr module (see chapter 1 1*Representation of Primary Sources* in the *Guidelines*) should be compatible with EVT, or made compatible with minor changes; on the image side, several features such as a magnifying lens, a general zoom, image-text linking and more are already available. For the future we aim at taking the Critical Apparatus module into consideration, which would imply creating a separate XSLT style sheet to complement the two existing ones, and at making it easier to configure the whole system, possibly by means of a GUI tool. Search functionality will be entrusted to a native XML database such as eXist.

EVT is built on open and standard web technologies, such as HTML, CSS and Javascript, to ensure that it will be working on all the most recent web browsers, and for as long as possible on the World Wide Web itself: specific features, such as the magnifying lens, are entrusted to jQuery plugins, again chosen among the open source, best supported ones to reduce the risk of future incompatibilities; the general architecture of the software, in any case, is modular, so that any component which may cause trouble or turn out to be not completely up to the task can be replaced easily. The project is nearing an alpha release (v. 0.2.0) on Sourceforge, and already offers all the tools listed above, with the exception of a search engine (expected to be implemented in v. 0.3.0).

Bibliography

Editions and digital facsimiles

- Biblioteca Apostolica Vaticana. http://www.vaticanlibrary.va/home.php?pag=mss_digitalizzati (accessed on March 2013).
- Codex Sinaiticus. http://www.codex-sinaiticus.net/en/manuscript.aspx (accessed on March 2013).

- e-codices. http://www.e-codices.unifr.ch/ (accessed on March 2013).
- e-sequence. http://www.e-sequence.eu/de/digital-edition (accessed on March 2013).
- Foys, Martin K. 2003. The Bayeux Tapestry: Digital edition [CD-ROM]. Leicester: SDE.
- Kiernan, Kevin S. 2011. Electronic Beowulf [CD-ROM]. Third edition. London: British Library.
- Malory Project. http://www.maloryproject.com/image_viewer.php?gallery_id=7&image_id=11&pos=1 (accessed on March 2013).
- Muir, Bernard James. 2004a. The Exeter anthology of Old English poetry: An edition of Exeter Dean and Chapter MS 3501 [CD-ROM]. Revised second edition. Exeter: Exeter University Press.
- Online Froissart. http://www.hrionline.ac.uk/onlinefroissart/ (accessed on March 2013).
- Samuel Beckett Digital Manuscript Project. http://www.beckettarchive.org/demo/ (accessed on March 2013).
- Stolz, Michael. 2003. Die St. Galler Epenhandschrift: Parzival, Nibelungenlied und Klage, Karl, Willehalm. Faksimile des Codex 857 der Stiftsbibliothek St. Gallen und zugehöriger Fragmente. CD-ROM mit einem Begleitheft. Hg. von der Stiftsbibliothek St. Gallen und dem Basler Parzival-Projekt (Codices Electronici Sangallenses 1).
- The Dead Sea Scrolls. http://www.deadseascrolls.org.il/ (accessed on March 2013).
- Vercelli Book Digitale. http://vbd.humnet.unipi.it/ (accessed on March 2013).

Software tools

- DFG Viewer. http://dfg-viewer.de/en/regarding-the-project/ (accessed on March 2013).
- DM Tools. http://dm.drew.edu/dmproject/ (accessed on March 2013).

- Scalable Architecture for Digital Editions. http://www.bbaw.de/telota/projekte/digitale-editionen/sade/ (accessed on March 2013).
- TEI Boilerplate. http://teiboilerplate.org/ (accessed on March 2013).
- TEICHI. http://www.teichi.org/ (accessed on March 2013).
- The TEIViewer project. http://teiviewer.org/ (accessed on March 2013).

Essays and reference

- Burnard, L., K.O.B. O'Keeffe, and J. Unsworth. 2006. Electronic textual editing. New York: Modern Language Association of America.
- Buzzetti, Dino. 2009. "Digital Editions and Text Processing". In Text Editing, Print, and the Digital World. Ed. Marilyn Deegan and Kathryn Sutherland, 45–62. Digital Research in the Arts and Humanities. Aldershot: Ashgate. http://137.204.176.111/dbuzzetti/pubblicazioni/kcl.pdf.
- Foys, Martin K., and Shannon Bradshaw. 2011. "Developing Digital Mappaemundi: An Agile Mode for Annotating Medieval Maps". Digital Medievalist n. 7. http://www.digitalmedievalist.org/journal/7/foys/ (accessed on March 2013).
- Landow, George P. 1997. Hypertext 2.0: The convergence of contemporary critical theory and technology. Baltimore: Johns Hopkins University Press.
- O'Donnell, Daniel Paul. 2005a. Cædmon's Hymn: A multimedia study, archive and edition. Society for early English and Norse electronic texts A.7. Cambridge and Rochester: D.S. Brewer in association with SEENET and the Medieval Academy.
- O'Donnell, Daniel Paul. 2005b. "O Captain! My Captain! Using technology to guide readers through an electronic edition." Heroic Age 8. http://www.mun.ca/mst/heroicage/issues/8/em.html (accessed on March 2013).
- O'Donnell, Daniel Paul. 2007. "Disciplinary impact and technological obsolescence in digital medieval

studies". In A companion to digital literary studies. Ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell. 65-81. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-4-2 (accessed on March 2013).

- Price, Kenneth M. 2008. Electronic Scholarly Editions». In A Companion to Digital Literary Studies. Ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell.
- Robinson, Peter. 2004. "Where We Are with Electronic Scholarly Editions, and Where We Want to Be". http://computerphilologie.uni-muenchen.de/jg03/robinson.html (accessed on March 2013).
- Rosselli Del Turco, Roberto. 2006. 'La digitalizzazione di testi letterari di area germanica: problemi e proposte'. Atti del Seminario internazionale 'Digital philology and medieval texts' (Arezzo, 19 – 21 Gennaio 2006), Firenze: Sismel.
- Rosselli Del Turco, Roberto. 2011. 'After the editing is done: designing a Graphic User Interface for Digital Editions.' Digital Medievalist Journal vol. 7. http://www.digitalmedievalist.org/journal/7/rosselliDelTurco/ (accessed on March 2013).
- TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. V. P5 (31 January 2013). http://www.tei-c.org/P5/.

Use of TEI in the Wolfenbuettel Digital Library (WDB)

Schaßan, Torsten; Steyer, Timo; Maus, David

This poster will present the use of TEI in the Wolfenbuettel Digital Library (WDB), housed by the Herzog August Bibliothek (HAB), and present the

ODDs applied, the ways of creation, processing models, workflows, and the appearance of TEI data in various contexts.

The WDB, that had been a publication platform for digitised cultural heritage materials (as images) in the first place is about to be transformed into a general publication platform for complex digital objects such as digital editions, combining images, full-texts of digitised (and OCRed) prints, and additional data on those digitised materials such as descriptions and structural metadata.

TEI plays an important role in this context as it is created and used in the WDB in various ways:

- as born digital format, e.g. during manuscript description and for digital editions;
- as automatically generated data during digitisation and OCR;
- as result of transformations from various sources, including conversions from PDF, InDesign, Word; the resulting data is used as publication format to "populate" the WDB;
- as storage format, boiled down to a standard encoding ("base format");
- as export format, especially towards repositories such as Europeana.

Data creation

The HAB is partner in various projects that produce TEI data in different ways:

- There are inhouse manuscript cataloguing projects that encode the descriptions directly in TEI, using the ODD and all materials provided by the previous MASTER and Europeana Regia projects. (cf. http://diglib.hab.de/rules/documentation/) Other inhouse projects prepare digital editions, again directly in TEI for the use and publication. The library has created a working group to set standards of encoding for both kinds of materials and helps respectively oversees the creation (and publication) of that data.

- With the WDB becoming more and more visible to others the library faces a rising number of requests to house externally prepared digital editions. Some base standards have to be set to match the needs of those externally prepared editions and their requests to be published within the WDB.
- The conversion of formerly printed text into a digital, structured full-text is a work more and more common. Within the library the modern works published by the library itself are object to this conversion as well as the OCR of historical prints mainly from the 17th century. The resulting texts need to have a common basic encoding. To set the level of this base encoding is addressed by the project AEDit. (cf. http://diglib.hab.de/?link=029)

Transformation

Data comes in all forms to the HAB: As Word, InDesign or PDF files, LaTeX encoded, XML in various flavours, and often as well as TEI files. Problems with TEI files are that the encoding has various flavours and encoding different depths. From these input formats conversions have to be organised into a harmonised TEI format.

Publication

All TEI data are used for publishing. In the scope of the WDB, XML data are exposed both as result of XSLT transformations into HTML and as source data that can be downloaded. In the case of manuscript descriptions HAB runs a manuscript database that is implemented using eXist. Additionally, eXist serves as search engine for the WDB.
The poster will also address interchange issues such as the use of TEI in combination with METS, and the mapping to and export towards ESE/ EDM.

Discovery, and Dissemination

The creation of digital editions and digital modelling of various dataformats, semantic searches and the visualisation of data are issues that touch on basic problems common to the diverse disciplines within digital humanities. The HAB currently runs the project "Digital Humanities" which will analyse cataloguing and indexing projects that rely on metadata

and explore how current standards and ontologies can be used for modelling central entities such as persons, corporate bodies and places. If necessary, such standards will be customized, developed further and applied as test cases within some current HAB projects. The focus will rest on normalizing data in order to allow an exchange between the various projects of the partners involved and enable an integration into existing or future search engines.
All data that is available via the HABs OAI (http://dbs.hab.de/oai/wdb/) is available under CC-BY-SA license. (cf. http://diglib.hab.de/copyright.html) All data produced inhouse is exposed in the WDB under the same license. Data produced by others and only published via WDB may be subject to other rights declarations.

The major issues of the poster will be both the transformations of TEI into a base format that can be easily used within the WDB as well as the question which TEI it exactly is that can be used this way.

Bibliography

- Stäcker, Thomas: Creating the Knowledge Site - elektronische Editionen als Aufgabe einer Forschungsbibliothek. In: Digitale Edition und Forschungsbibliothek. Ed. Christiane Fritze et al. Wiesbaden 2011, p. 107- 126 (Bibliothek und Wissenschaft, 44)

The Bibliotheca legum project

Schulz, Daniela Monika

Medieval law is a research field of interest to historians, medievalists as well as legal scholars. Especially regarding the past it is often quite difficult to determine what applicable law actually was. The "Bibliotheca legum regni Francorum manuscripta" project ("Bl") aspires to do so with a focus on the legal knowledge that was prevalent in the Francia. All

"leges" (secular law texts) that were copied during the Carolingian period are incorporated.

The website provides an introductory text to each "lex" including reading recommendations, as well as short descriptions of all codices containing these texts. Information on repository, origin and history of the manuscript, contents as well as bibliographical references etc. are given. At the moment there are 273 short descriptions available.

The aim of the Bl is to take up the current state of research and also the research history as complete as possible. Therefore a lot of effort was put into gathering this information. All prior studies concerning single manuscripts as well as several editions of the law texts were surveyed. For each manuscript, age determinations and assumptions about its origin carried out by the different describers are recorded. Thus the features of the various print editions are transposed into the electronic version.

Originally the information was gathered in a MS-Word table, since it was prepared for internal use only. This had certain impacts on the procedural method. The idea to make the data publicly available in a digital form emerged in summer 2012, so the Bl is in its initial year of development. It is work-in-progress and not officially launched yet. Although not all functionalities as well as information are available by now, it is accessible on the web. This was a willful decision to enable the public to pursue the genesis and the development of the project.

The Bl heavily relies on existing resources. With regard to the needs of academic research, it gathers all digital images available of the respective manuscript testimonies (e.g. from "Europeana", "Gallica") as well as catalogue information (e.g. "Manuscripta Medievalia"). Therefore the Bl can be seen as a meta-catalogue and gateway to further resources. With kind permission of the "Monumenta Germaniae Historica" (MGH) it was possible to also integrate the complete text of the "Bibliotheca capitularium regum Francorum manuscripta. Überlieferung und Traditionszusammenhang der fränkischen Herrschererlasse" by Hubert Mordek (Munich 1995), which is the most comprehensive work on codices from the respective period. It is not only downloadable in its totality of more than 1000 pages as a PDF, but also as compilations of pages regarding single manuscripts that have been described by him.

The encoding is carried out according to the TEI P5 standard. People and places are tagged and enhanced according to authority files such as VIAF or TGN to enable identification. Wordpress is used as a CMS for data management and to provide basic functions. While this platform is very common in the World Wide Web, it is not widely adopted for Digital Humanities' projects working with XML data. The XSLT processing of the XML files within Wordpress as well as certain other features (multilingualism, viewers etc.) are realized by plugins. The Bl is published under Creative Commons licence. XML source files are provided for all manuscripts and are freely available for download.

Features

- The BL is a multi-language site with interface and general information in German and English.
- Manuscript descriptions can be reached via multiple browsing accesses (shelfmark, leges contained, date of origin, place of origin).
- Full text and faceted search are included.
- All resources within the Bl as well as the external ones are connected via inter-/hyperlinking.
- Information is given on different levels to make this platform a useful tool for scholars, students and the interested public audience.
- A comprehensive bibliography on the subject and indices on people, places as well as repositories facilitate further orientation and provide contextualization.
- Each manuscript description is available as XML download.
- Some prior studies and editions are integrated within a viewer and are also available as PDF downloads.
- A blog (German / English) informs about the current state of development and related topics.

The presentation might be of interest to all those working in projects that evolve under similar conditions, namely

- relatively small workforce (3),

- no funding,
- lack of previous experiences in setting up a DH project from the scratch, and
- absence of a technical partner or "real" programming / web developing skills.

The poster will present the lessons learnt during the initial year of development. Emphasis will be on the use of TEI and TEI connected tools, the difficulties encountered and the compromises made. Also a comprehensive evaluation of Wordpress as a CMS within the TEI/XML context is included.

Staff

Prof. Dr. Karl Ubl, Chair of Medieval History, Cologne University (Project Lead)
Dominik Trump (Data aggregation and text encoding)
Daniela Schulz (Technical Lead)

References

- Hubert Mordek, Bibliotheca capitularium regum Francorum manuscripta. Überlieferung und Traditionszusammenhang der fränkischen Herrschererlasse (MGH Hilfsmittel 15), München 1995.
- http://www.leges.uni-koeln.de
- http://www.tei-c.org
- http://www.europeana.eu/
- http://gallica.bnf.fr/
- http://www.manuscripta-mediaevalia.de
- http://www.mgh.de/
- http://www.wordpress.com
- http://viaf.org/
- http://www.getty.edu/research/tools/vocabularies/tgn/
- http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/handschriften.pdf

Digital edition, indexation of an estate, collaborations and data exchange – August Boeckh online

Seifert, Sabine

The August Boeckh project is one of the major research initiatives of the junior research group "Berlin intellectuals 1800–1830", led by Dr. Anne Baillot at Humboldt University Berlin. The project focuses on August Boeckh's (1785–1867) manuscripts, who was one of the most important German classical philologists and a central figure in nineteenth-century Berlin. The Boeckh project can be seen as an example of collaboration between institutions, and of developing strategies to link the (meta-)data from libraries and archives with research results. We cooperate with archives and libraries such as the State Library Berlin and Humboldt University Library. Thus, the project is designed to be broader in scope with many connecting factors, and suitable for data exchange.

The key aspects considered for edition and interpretation are (a) the indexing of Boeckh's literary estate for the August Boeckh Online Platform; (b) the edition of selected letters and reports from this estate as part of the digital edition "Letters and texts. Intellectual Berlin around 1800"; (c) the edition of Boeckh's manuscript for his lecture "Encyklopädie und Methodologie der philologischen Wissenschaften", a major work in the history of the classics; and (d) a virtual reconstruction of Boeckh's personal library consisting of approximately 12,000 books. All these sub-projects help to reconstruct Boeckh's horizon of knowledge and gain insight into his scholarly work und understanding. With this poster, I want to concentrate on the first two aspects.

The August Boeckh Online Platform presents Boeckh's extensive literary estate in a systematic overview which has been an acknowledged desideratum.[47] The first step was the detailed indexing of each individual manuscript and letter by Boeckh in several Berlin archives and libraries, with a short summary of the content. In addition to these approximately 1500 entries in XML/TEI P5 format, up to 900 entries related to Boeckh

[47] Baillot, Anne, "August Boeckh – Nachlassprojekt". [http://tei.ibi.hu-berlin.de/boeckh/] The Platform will be publicly available in August 2013. Login: berlin, password: heidelberg.

from the Kalliope manuscript database in XML were imported.[48] At this step, we encountered a problem of disparities in the level of indexing because the Kalliope entries are often based on boxes instead of single documents (one box of e.g. 50 letters versus one letter). Our aim is to complete this information to have rich metadata on every single document in Boeckh's literary estate. Then, the data will be submitted to Kalliope, so that Kalliope benefits from our research results. At a later stage, the same process will involve data exchange with the Humboldt University Library for reconstructing Boeckh's library.

The project is also overseeing the publication of selected letters and reports from the estate concerning Boeckh's activities at the Berlin university, especially his philological seminar.[49] These previously unpublished documents shed light on the development of the university and research policy in nineteenth-century Prussia, and are part of the digital edition "Letters and texts. Intellectual Berlin around 1800". The edition centres around the main research question of intellectual networks in the Prussian capital city Berlin at the beginning of the nineteenth century, and publishes letters and work manuscripts by a selection of several authors.[50] The connection with the Boeckh Platform is ensured by a specific XML/TEI P5 schema that is documented in our encoding guidelines.[51] The indices play a central role and contain information from our several projects and constantly interlink them. As with the Boeckh Online Platform, our goal is to exchange data and link with other projects

[48] http://kalliope-portal.de/

[49] Seifert, Sabine (ed.), "August Boeckh", in: Anne Baillot (ed.), "Letters and texts. Intellectual Berlin around 1800", Humboldt University Berlin [in preparation, 2013] [http://tei.ibi.hu-berlin.de/berliner-intellektuelle/author.pl?ref=p0178]. On Boeckh's founding of and directing the philological seminar, see Sabine Seifert, "August Boeckh und die Gründung des Berliner philologischen Seminars. Wissenschaftlerausbildung und Beziehungen zum Ministerium", in: Christiane Hackel, Sabine Seifert (eds.), August Boeckh. Philologie, Hermeneutik und Wissenschaftspolitik (Berlin, 2013), pp. 159–178.

[50] For an introduction to this digital edition as well as its use in teaching, see Anne Baillot, and Sabine Seifert, "The Project 'Berlin Intellectuals 1800–1830' between Research and Teaching", in: Journal of the Text Encoding Initiative [Online] Issue 4 (March 2013) [http://jtei.revues.org/707; DOI : 10.4000/jtei.707].

[51] http://tei.ibi.hu-berlin.de/berliner-intellektuelle/encoding-guidelines.pdf

and institutions. Thus, the data architecture of the digital edition needs to be detailed as well as open.

The text of the manuscripts is presented in a diplomatic transcription and in an edited version,[52] both generated from the same TEI P5 file. The encoding of letters posed some problems, as there is no letter-specific TEI module yet.[53] In these cases, we consulted the SIG Correspondence and other digital editions, such as the Carl Maria von Weber – Collected Works.[54] Viewing our transcription, the user can compare it with a facsimile of the manuscript as well as with the XML file containing the metadata and the encoding. The XML files are published under a CC-BY licence that they can be re-used and enriched for further research. In the edition and the Boeckh project, authority files are used whenever possible[55] including the GND for the identification of persons (Integrated Authority File,[56] via entries of the GND number in our index of persons), in collaboration with the Person Data Repository at the Berlin-Brandenburg Academy of Sciences;[57] the use of ISO-Codes; persistent URLs (the collaboration with libraries is especially important in this regard because they are probably the only ones who can provide these URLs); individual IDs for each XML/TEI document, etc. In order to answer our main research questions of how intellectual networks were established, how transfer of knowledge took place and books were read or produced, and to reconstruct – and visualize – the dynamics of group

[52] See Sperberg-McQueen on the fact how technical possibilities and the mutability of digital presentation influence editing as well as editorial theory, C. M. Sperberg-McQueen, "How to teach your edition how to swim", in: LLC 24,1 (2009), pp. 27–39, esp. pp. 31–33 [DOI: 10.1093/llc/fqn034].

[53] On the treatment of correspondence in scholarly editing in general and on the problems of encoding correspondence in TEI, see Edward Vanhoutte, Ron Van den Branden, "Describing, transcribing, encoding, and editing modern correspondence material: a textbase approach", in: LLC 24,1 (2009), pp. 77–98, esp. pp.82–90 [DOI: 10.1093/llc/fqn035].

[54] http://www.weber-gesamtausgabe.de

[55] On the concepts of authority files and their use in scholarly editions, see Peter Stadler, "Normdateien in der Edition", in: editio 26 (2012), pp. 174–183 [DOI: 10.1515/editio-2012-0013].

[56] http://www.dnb.de/EN/Standardisierung/GND/gnd.html

[57] http://pdr.bbaw.de/english

relationships, there is a mark-up for people, places, works (e.g. books, articles), and groups/organisations. Via these aforementioned indices the user can search in the edition's other corpora that cite people, works etc. also cited in the edited Boeckh manuscripts. When used in connection with the Boeckh Online Platform, the researchable context becomes even more comprehensive. On both front ends, search results are shown for the edition as well as the platform and, thus, the manifold connections between the several corpora in the edition (i. e. the manuscripts) and the Platform (i. e. metadata on these and other manuscripts) are made manifest.

In this poster, I want to present the August Boeckh Online Platform and its connection to the digital edition "Letters and texts. Intellectual Berlin around 1800" in the many aspects offered by the manuscripts. I will demonstrate the workflow of the cooperations with the libraries and the wide range of documents that can be linked to the edition with the help of these connections. Furthermore, I will develop one example (the philological seminar) to show how research can benefit from such an approach.

Bibliography

- Baillot, Anne, "August Boeckh – Nachlassprojekt" [http://tei.ibi.hu-berlin.de/boeckh].
- Baillot, Anne; Seifert, Sabine, "The Project 'Berlin Intellectuals 1800–1830' between Research and Teaching", in: Journal of the Text Encoding Initiative [Online] Issue 4 (March 2013) [http://jtei.revues.org/707 ; DOI : 10.4000/jtei.707].
- Seifert, Sabine (ed.), "August Boeckh" [http://tei.ibi.hu-berlin.de/berliner-intellektuelle/author.pl?ref=p0178], in: Anne Baillot (ed.), "Letters and texts. Intellectual Berlin around 1800", Humboldt University Berlin (Berlin, 2013) [http://tei.ibi.hu-berlin.de/berliner-intellektuelle/?language=en].
- Seifert, Sabine, "August Boeckh und die Gründung des Berliner philologischen Seminars. Wissenschaftlerausbildung und Beziehungen zum Ministerium", in: Hackel, Christiane; Seifert, Sabine (eds.), August Boeckh. Philologie, Hermeneutik und Wissenschaftspolitik (Berlin, 2013), pp. 159–178.

- Sperberg-McQueen, C. M., “How to teach your edition how to swim”, in: LLC 24,1 (2009), pp. 27–39 [DOI: 10.1093/llc/fqn034].
- Stadler, Peter, „Normdateien in der Edition“, in: editio 26 (2012), pp. 174–183 [DOI: 10.1515/editio-2012-0013].
- Vanhoutte, Edward; Branden, Ron Van den, “Describing, transcribing, encoding, and editing modern correspondence material: a textbase approach”, in: LLC 24,1 (2009), pp. 77–98 [DOI: 10.1093/llc/fqn035].

'Spectators': Digital Edition as a tool for Literary Studies

Semlak, Martina; Stigler, Johannes

The proposed poster presents the digital edition of about 30 Romanic moral weeklies ('Spectators') as an example of how the TEI can be used for a project which has to deal with complex and overlapping text structures, a large corpus of texts and a data creation environment involving staff with no special training in XML and TEI.
'Spectators' are a journalistic genre which had its origins at the beginning of the 18th century in England and spread out all over Europe. It became an important feature of Enlightenment and distributed ethical values and moral concepts to a broad, urban readership. The objective of this digital edition (http://gams.uni-graz.at/mws) is both editing the prominent Romanic weeklies and analysing the texts on narratological and thematic levels. Currently 1300 Spanish, French and Italian texts are provided. The collection is continuously expanded. The project has been realized as a co-operation by the department for Romance Studies and the Center for Information Modeling at the University of Graz.
A characteristic feature of the text genre of 'Moral Weeklies' or 'Spectators' are interruptions of the text flow and overlays of narrative structures

that result from the change of actors and real and fictional dimensions. One goal was a faithful reproduction of the individual issues regarding text-logical units such as headings, paragraphs or quotes. An additional demand was the enrichment of the material by adding research results from the analysis of display planes and narrative forms such as dialogue, letter or dream narrative etc. inside the texts. These considerations were formalized in a data model corresponding to the requirements for an explication of the linguistic and narrative structures, based on TEI P5 XML.

A particular editorial challenge of this digital edition are the overlapping structures resulting from the text-logical units and narrative forms. To solve this problem, the TEI provides different strategies. In this project we decided on using boundary marking with empty elements to mark the starting and ending points of levels of interpretation and narrative forms.

For the implementation of the digital edition of the 'Spectators', an exemplary work flow was developed. In addition to the assessment of the material and the survey of the project objectives, this work flow includes a data acquisition scenario which supports the digital compilation and semantic analysis of the research data by the scholars: Based on the data model, a document template for a standard text processing program is created, which includes macros to transform the input into a TEI document.

A webbased Java client allows for the upload of documents into the repository, the Geisteswissenschaftliches Asset Management System (GAMS), which meets the requirements of the OAIS reference model. Based on the open source project FEDORA, this object-oriented digital archive offers the individual design and flexible adaptation of content types (‘content models’) tailored to the type and scope of the source material and specific research interests. A 'content model' describes the structural components of a digital object, essentially consisting of a persistent identifier, metadata, content and disseminators.

More than 1300 texts from some 30 French, Italian and Spanish 'Spectators' have already been published in their original language using the methods outlined above. The data is available under a Creative Commons license. Moreover, the objects are integrated into the European

search portal Europeana (http://www.europeana.eu). The user interface to the collection is multilingual.

Bibliography

- Ertler, Klaus-Dieter (2012): "Moralische Wochenschriften", in: Leibniz-Institut für Europäische Geschichte (IEG): Europäische Geschichte Online (EGO). Mainz 2012. http://www.ieg-ego.eu/ertlerk-2012-de
- Hofmeister, Wernfried/Stigler, Hubert (2010): "Edition als Interface. Möglichkeiten der Semantisierung und Kontextualisierung von domänenspezifischem Fachwissen in einem Digitalen Archiv am Beispiel der XML-basierten 'Augenfassung' zur Hugo von Montfort-Edition", in: Nutt-Kofoth, Rüdiger / Plachta, Bodo / Woesler, Winfried: editio. Internationales Jahrbuch für Editionswissenschaft. Berlin, New York: Walter de Gruyter, 79–95.
- Lagoze, Carl/Payette, Sandy/ Shin, Edwin/Wilper, Chris (2006): "Fedora. An Architecture for Complex Objects and their Relationships". http://arxiv.org/ftp/cs/papers/0501/0501012.pdf
- Vasold, Gunter (2013): "Progressive Editionen als multidimensionale Informationsräume", in: Ambrosio, Antonella / Barret, Sébastien / Vogeler, Georg: Digitale Diplomatik 2013. Tools for the Digital Diplomatist, Köln: Böhlau, in print.

Laundry Lists and Boarding Records: challenges in encoding "women's work"

Tomasek, Kathryn; Bauman, Syd

Introduction

In 'Encoding Financial Records for Historical Research' presented at this conference last year in Texas and slated for publication in an upcoming issue of the Journal of the Text Encoding Initiative, we noted a shortcoming of current TEI encoding methods for representing services, as opposed to commodities, when being transfered or traded: In many cases one of the 'items' being transferred is a service, not a commodity. Our current system, being based on the TEI `<measure>` element, seems a clumsy way to handle this. For example, measure unit="hours" quantity="2" commodity="babysitting" may be reasonable, but when the service being provided is recorded either by things *on* which it is performed or the people *for* whom it is provided, rather than the amount *of the service* that is provided, it becomes difficult to express formally using the current system. The 'transactionography' approach described in that paper relies on the TEI `<measure>` element to record the *what* of a transfer. (The *when* is recorded using the TEI att.datable attributes.) Many historical financial records, however, include or are even primarily about the exchange of money for services (e.g., laundering, room and board, or domestic service). Since these services were more usually performed by women and often recorded by women, study of these types of HFRs is of particular interest to practitioners of women's history.

Sample Problems

The quintessential example of this problem occurs when trying to encode a 'laundry list'. Such lists include a set of items of clothing and prices. But the price is not for *purchasing* the associated item of clothing, but for *laundering* it (which is often not explicitly stated).

While one might claim that the work of laundering is implied by the genre 'laundry list,' such generic information must be recorded somehow in order to be machine-readable. If we use the `<list>` element, the **`@type`**

and **@subtype** attributes could be used to express that the costs listed are for laundering, not purchasing, but there is no agreed upon vocabulary with which to express this, and it may not generalize well to other services. Many examples of such laundry lists are extant, and they can potentially provide information not only about period clothing and the habits of wearers, but also about the comparative value of laundering services in different regions and periods, and perhaps (with sufficient contextual information) about the relative cost (and therefore value) of the work of laundering in such various contexts as an individual laundress subcontracting with the keeper of a boarding house, an insitutional laundry as a department of a hospital or hotel, or an industrial laundry serving individual or institutional clients.

In the case we will show in the poster, an individual laundress subcontracted with a boardinghouse keeper to perform the service of laundering clothing and household linens for people who also rented rooms and purchased meals at the boarding house. The laundry lists make up one set of documents that record exchanges of services for cash. They are supplemented by small notebooks in which the boardinghouse keeper tracked charges for food and such other necessities as candles and soap, as well as weekly payments for room and board. The boarder also kept receipts as her own record of the payments.

One Possible Solution

In our 'transactionography' we have heretofore used the TEI `<measure>` element, with its **@quantity**, **@unit**, and **@commodity** attributes, to represent that which is transferred from one person or account to another in a transaction. But in the laundry list case, the work performed by the laundress is not a "commodity" but a "service," the service for which the boarder paid the boardinghouse keeper in this transaction. However, using the `<measure>` element with existing attributes leads to markup that fails to distinguish the purchase of a garment from paying for the service of laundering it. One possible solution is to add a new attribute, **@service.** Thus for instance, a line from a laundry list might be marked up as follows.

```
<hfr:transaction>
<hfr:transfer fra="people.xml#fearn" til="people.xml#EW">
<measure quantity="2" unit="count" commodity="skirt"
service="laundering">2 wool skirts</measure>
</hfr:transfer>
```

```
<hfr:transfer fra="people.xml#EW" til="people.xml#fearn">
<measure quantity="6" unit="pence" commodity="currency"
>6</measure>
</hfr:transfer>
</hfr:transaction>
```

This solution seems to have broad application. E.g.:

- Framing: measure quantity="15" unit="count" commodity="8x10 color glossies" service="framing"
- Shoe shining: measure quantity="2" unit="count" commodity="shoe" service="shining"
- XSLT programming: measure quantity="18" unit="hours" service="programming"

We will not be surprised, however, if there are cases it does not handle well.

A Broader Problem?

The issues presented by the laundry list example may be representative of a larger problem, that of indirect reference. Indirect reference was described in 2008 by the Women Writers Project. This phenomenon occurs when an author refers to one entity by naming another. In the WWP's case a person is referred via the name of another person, character, or figure. E.g., the headline of a 2007-05-31 article in the Toronto Star, Terminator gunning to save lives, refers to then governor of California Arnold Schwarzenegger indirectly through a reference to a character he played in a well-known film. The WWP solution addresses this specific use-case, `<persName>`:

… to represent the special nature of metaphorical or figurative references.

… For this …, the WWP has created a custom attribute for `<persName>`, **`@wwp:metaRef`**. For practical purposes, **`@wwp:metaRef`** functions exactly like **`@ref`**; where **`@ref`** points to the unique **`@xml:id`** for the actual reference, however, **`@wwp:metaRef`** points to the **`@xml:id`** of the person being indirectly or figuratively referenced. For example —

Source text:

```
Come all ye tender Nymphs and sighing Swains,
Hear how our Thyrsis, Daphnis death complains
```

Encoded text:

```
<l>Come all ye tender Nymphs and sighing Swains,</l>
<l>Hear how our <persName ref="personography.xml#thyrsis.auc"
wwp:metaRef="personography.xml#jfroud.jke">Thyrsis</persName>,
<persName ref="personography.xml#daphnis.tvc"
wwp:metaRef="personography.xml#tcreech.zxz">
```

```
Daphnis</persName> death complains</l>
```

It occurs to us that that these cases may not be very different. In the laundry list example, the work of laundering a skirt is referred to by reference to the skirt itself. In the Toronto Star example Arnold Schwarzenegger is referred to by reference to the character he played. Each is a case of indirect reference. It is interesting to contemplate a generic TEI mechanism for indirect reference that would handle both cases.

Conclusion

In this poster presentation we hope to frame the problem of encoding services within historical financial records, present at least one possible solution, and solicit input from the attendees of the TEI conference about the utility of our proposed solution, and about other possible encoding methodologies to solve this shortcoming. One goal is to come up with a methodology that might apply to other cases of what might be called *indirect reference*.

Bibliography

- Tomasek, Katheryn Bauman, Syd 'Encoding Financial Records for Historical Research' *Journal of the Text Encoding Initiative* forthcoming
- Melson, John Flanders, Julia 'Not Just One of Your Holiday Games: Names and Name Encoding in the Women Writers Project Textbase' http://www.wwp.brown.edu/research/publications/reports/neh_2008/WWP_Names_White_Paper.pdfhttp://www.wwp.brown.edu/research/publications/reports/neh_2008/WWP_Names_White_Paper.pdf

TEI/XML Editing for Everyone's Needs

Wiegand, Frank

Project

The DFG-funded project Deutsches Textarchiv (DTA) started in 2007 and is located at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). Its goal is to digitize a large cross-section of German texts published between 1600 to 1900. The DTA presents almost exclusively the first editions of the respective works. Currently, the DTA core corpus consists of more than 840 texts, which were transcribed mostly by non-native speakers using the double keying method. In addition, the DTA hosts more than 520 further texts which were imported using the extension module DTAE. In total, the corpus consists of more than 380,000 text pages (June 2013).

The DTA provides linguistic applications for its corpus, i. e. serialization of tokens, lemmatization, POS tagging, lemma based search, and phonetic search based on rewrite rules for historic spelling. Each text in the DTA is encoded using the DTA base format (DTABf), a strict subset of TEI P5. The markup describes text structures (headlines, paragraphs, speakers, poem lines, index items etc.), as well as the physical layout of the text.

Quality assurance for all texts within the DTA corpora takes place within the quality assurance platform DTAQ. In DTAQ, texts may be proofread page by page in comparison to their source images. This way errors can be detected which may have occurred during the transcription and annotation process.

Problem Statement

DTAQ is running since March 2011, and a lot of tools were developed which allow for various kinds of annotations to the digitized texts. DTAQ users should be enabled not only to review texts but also to correct erroneous transcriptions and annotations, or add new annotations. Within DTAQ, each text is presented page by page alongside its source images in various formats: XML, HTML, plain text etc. To produce this kind of view, the original TEI P5 documents are splitted into several single page

documents. This process is reversible, so modified single page documents can be reinserted losslessly into the original TEI document. Based on this page-oriented view, DTAQ provides several ways to change documents on the transcription or annotation level.

We differentiate between several kinds of changes and user experience levels:

- Changes to the text base (i. e. the plain transcribed text without any kind of markup).
- Annotation of single tokens or groups of tokens, e. g. named entity annotation, annotation of printing errors etc.
- Editing of attribute values in existing XML elements, e. g. the values of **`@ref`** in `<persName>` elements to provide links to authority files.
- Editing of basic XML structures, e. g. adding quotation markup in citations (`<cit>`/`<quote>`/`<bibl>`).
- Editing of more complex XML structures, e. g. restructuring of paragraphs or even chapters.

For some of these kinds of changes users may not even have to bother with XML markup, other changes require a deeper look into the complete XML document, e. g. if they occur across page breaks, or could produce overlapping hierarchies.

Even though there is a comprehensive documentation available for the DTABf, less experienced users (especially those with little if any previous knowledge of the XML standard) would have to spend significant amounts of time to learn how to properly apply changes to the TEI documents on the level of transcription or annotation.

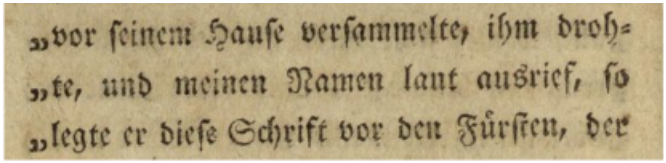
In addition, each change must be tracked within a revision history system to see (and moderate), which user changed the texts within the DTA repository.

Various Editing and Annotation Possibilities

To make easy changes easy and hard things possible, we provide several ways for users to deal with the digitized texts:

Instant WYSIWYG Editor

Simple changes, like fixing transcription errors, may be carried out directly within the rendered HTML version of a document page, using the `@contenteditable="true"` attribute (cf. http://www.w3.org/TR/2010/WD-html5-20100624/editing.html#contenteditable) which is available within all modern browsers. This technique allows for real WYSIWYG (what you see is what you get), because it makes the generated HTML editable within the rendered view. The modified text node is sent back to the repository, where it replaces the text node of the original TEI document. Users cannot produce invalid markup and don't even have to bother with angle brackets (cf. http://philomousos.blogspot.de/2011/01/i-will-never-not-ever-type-angle.html).



„vor ſeinem Hauſe verſammelte, ihm droh-
„te, und meinen Namen laut ausrief, ſo
„legte er dieſe Schrift vor den Fürſten, der

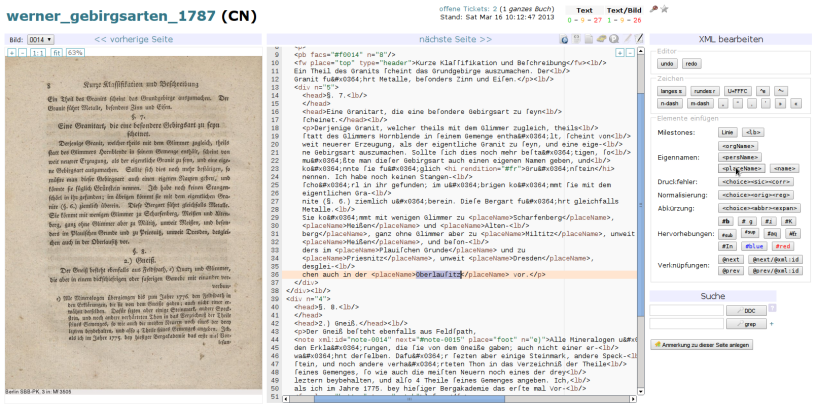
Simple Annotation Editor

To annotate simple phrases like named entities, no further knowledge of XML is needed. Just like in the correctors' view, where users can proofread pages, mark erroneous passages with their mouse, and report the errors via a ticketing system, named entities can be marked and labeled as `<persName>` or `<placeName>`, and additional data like references to an authority file can be provided using the **`@ref`** attribute.

XML Editor for Single Pages

For mid-size changes on single pages, we provide an online XML editor. This tool is based on the Ajax.org Cloud9 Editor (ace). The editor window displays the syntax-highlighted XML for the corresponding text page. In addition, we provide several

tools to support quick and efficient tagging (e. g. select a wrongly printed word like “errorneous”, press the “printing error” button, and an XML template like `<choice><sic>errorneous</sic><corr></corr></choice>` is inserted into the editor). The editor also provides validation against the DTABf schema (via AJAX requests).



DTA oXygen Framework

For larger changes, or even the beginning of new transcriptions from scratch, the DTA developed DTAoX, a framework for the widely used oXygen XML editor, which supports text editing in conformity with the DTABf within oXygen's author mode. A fine grained color scheme provides visualisations of different tagging levels (as well as of discrepancies with regard to the DTABf) to produce DTABf compatible TEI files. To apply changes to DTA documents, users have to download the whole TEI documents from the DTA repository, mark them as “locked” (to avoid conflicts with other changes), perform their intended changes, and upload the modified documents back into the repository.

Tracking Changes with git

Each time a change is submitted to the repository, the resulting document is validated against the DTA base format schema, and rejected, if the validation fails. Otherwise, the document gets an updated set of metadata (esp. with regard to timestamps and editor's responsibilities) and is committed to a git repository. We chose git, because—in contrast to other source control systems—git can deal with huge XML files adequately. Using a version control system of course is crucial, for every change needs to be reproducible and reversible, if necessary.

Availability

The DTABf documentation with a lot of illustrated examples from the DTA core corpus is freely available at the DTA website. RNG and ODD files are provided, as well as template files for starting a new transcription project.

DTAoX, the DTA oXygen framework is freely available for download under the LPGL license.

In its third project phase (application is currently under appraisal by the DFG), the DTA project will provide the DTAQ quality assurance framework for a wider audience and making it open source under the LGPL license.

Poster Presentation and Live Demonstration

The poster will provide a detailed insight into the various text editing modes the DTA provides. Visitors will be able to try out the respective tools by themselves at the live presentation desk.

References

[1] Ajax.org Cloud9 Editor: http://ace.ajax.org.

[2] Cayless, Hugh: I Will Never NOT EVER Type an Angle Bracket (or IWNNETAAB for short). In: Scriptio Continua, 2011-01-06. http://philomousos.blogspot.de/2011/01/i-will-never-not-ever-type-angle.html.

[3] Deutsches Textarchiv: Basis for a Reference Corpus for the New High German Language. http://www.deutschestextarchiv.de.

[4] DTA base format (DTABf): http://www.deutschestextarchiv.de/doku/basisformat.

[5] DTA oXygen framework (DTAoX): http://www.deutschestextarchiv.de/doku/software#dtaox.

[6] oXygen XML editor: http://www.oxygenxml.com.

[7] git (distributed version control system): http://git-scm.com.

[8] W3C: HTML5. A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft 24 June 2010. http://www.w3.org/TR/2010/WD-html5-20100624/spec.html (see also the latest editor draft of the HTML 5 specification: http://www.w3.org/html/wg/drafts/html/master/).

Tutorial and workshop

Perspectives on querying TEI-annotated data

Banski, Piotr; Kupietz, Marc; Witt, Andreas

The TEI provides mechanisms to richly annotate a variety of digital resources used in the Humanities. The typical way in which many Humanities scholars use annotations is as instructions for processing them for the purpose of visualisation or transformation into other formats. However, a major aim of TEI annotation is to enrich the data with the results of scholarly effort. It is therefore essential to be able to efficiently retrieve the various pieces of information in a structured way. This, in turn, requires accessible and user-friendly -- but at the same time reasonably powerful -- query languages.

Naturally, XQuery or XSLT provide access to all the information expressed in annotations. However, it should be borne in mind that, despite the warm feeling of power that good command of XQuery or XSLT offers to the researcher, not everyone is able to exploit their full capacity. Learning either of these Turing-complete programming languages requires an amount of time and devotion that not every scholar or student is able to allocate for this purpose. Like in the case of natural languages, one benefits greatly from long-time exposure and repetition – but these are conditions that characterise the tasks that face programmers or IT personnel rather than most literary scholars or students, who may greatly benefit from more specialized query languages which are at least one level of abstraction above XSLT or XQuery, and which offer user-friendliness instead of ultimate power and versatility.

The world of Digital Humanities – arguably the central focus of the TEI – has long ago expanded beyond simple working with electronic text in the word processor of the day. DH specialists gather, curate, and query various sorts of textual data, from plain text via semi-structured XML to records in relational databases. The nature of the objects of research varies as well: they come, among others, as single texts with sometimes very complex internal structure, bundles of base documents with hierarchies of annotations and all kinds of interrelationships among them, parallel multilingual data (e.g. original works and their translations) or scattered

prosopographic fragments. Much of that can nowadays be wrapped in a TEI envelope.

Given the above issues, it is natural to wonder whether the strategy typically advocated in the work of the TEI Council and often voiced on TEI-L – to stress that the TEI should best be handled by general-purpose XML-oriented tools (to which XQuery and XSLT belong) – should carry over to the task of retrieval from richly annotated data, especially if said retrieval is to be made available to an average scholar or student. Or, more precisely, whether it would be better to offer scholars and students a language tied more tightly to the TEI data model and whether it is possible for such a query language to address the entire TEI universum of objects in a uniform manner.

Within the last decade, a lot of effort to create efficient and user-friendly query systems has been undertaken within corpus linguistics, but the knowledge about them spreads very slowly outside this field. On the other hand, corpus linguists are often not aware of specific issues and needs of querying digital texts used outside linguistics.

Therefore, the workshop aims at building a common ground for the sharing of experiences among researchers dealing with various aspects and forms of TEI-annotated digital text. The presentations will address the impact of experiences of querying richly annotated linguistic corpora on other fields within Digital Humanities and discuss specific TEI-related problems when dealing with queries.

We would like to invite contributions addressing, but not limited to, the following range of issues:

- query languages and query environments;
- queries dealing with a variety of text objects in a variety of TEI-annotated structures;
- enhancement of user-friendliness by, e.g., hiding the potential complexity under a simple set of agreed symbols or by the use of a graphical user interface;
- a common query language to extend over the range of objects defined by the TEI data model.

This workshop is meant to bring together, on the one hand, corpus linguists and computer scientists, who will present their suggestions of

reflections on the possibility of creating a Corpus Query Lingua Franca for Humanists, and, on the other, TEI practitioners themselves, presenting both concrete tasks that combine textual and non-textual data in a novel manner, as well as theoretical challenges that a modern query system for Digital Humanists should tackle.
Workshome homepage is to be found at http://corpora.ids-mannheim.de/queryTEI.html

Use of EpiDoc markup and tools: publishing ancient source texts in TEI

Bodard, Gabriel; Baumann, Ryan; Cayless, Hugh; Roued-Cunliffe, Henriette

EpiDoc is a set of guidelines for encoding ancient source texts in TEI (originally developed for Greek and Roman epigraphy, but now much more diverse, see list of projects, below), including a recommended schema and ODD, a lively community of practice and an ecosystem of projects, tools and stylesheets for the interchange and exploitation of such texts. This tutorial will introduce participants to the principles and practices of EpiDoc encoding, which are largely based on the practice of encoding single-source documents and the ancient objects on which they are written, as well as some of the tools and other methods made available by the community for transforming, publishing, querying, exchanging and linking of encoded materials.
We expect participants to have basic familiarity with the principles of XML and TEI, and some understanding of epigraphic practice and the Leiden Conventions would be an advantage, but so long as there is willingness to learn fast the programme should be of interest to beginners as well. Students are welcome to bring their own texts to work with, but examples will be provided by the tutors.
Programme:

- Day 1: Getting data into EpiDoc
  Morning: Introduction to EpiDoc encoding, Leiden Conventions, and object description/history. Example texts will be offered, with opportunity to practice encoding in EpiDoc. Most examples will be in Greek or Latin, but knowledge of these languages is not essential to participation.
  Afternoon: Introduction to Papyrological Editor (papyri.info/editor), the principles of the Leiden+ shorthand and the SoSOL workflow management tool behind it. Opportunity to use "tags-free" editing interface and further encoding practice. Discussion of applicability of SoSOL to other projects (e.g. annotation functions added by Perseus Project) and other methods and principles for converting digital texts to EpiDoc. Discussion of ways to convert legacy data in databases or text documents to EpiDoc. Participants who have documents in other formats that they would like to convert to EpiDoc are invited to bring them.
- Day 2: Exploiting and converting EpiDoc texts
  Morning: Searching EpiDoc. We shall provide a walkthrough of setting up the eXist XML database, loading texts into it, and searching with XQuery, including setting up Apache Solr and indexing documents via XSLT. Students will have an opportunity to try setting up a webservice to access and search datasets.
  Afternoon: Publishing EpiDoc as Linked Data. Discussion of Linked Data principles and how these apply to setting up an infrastructure for publishing EpiDoc. Linking EpiDoc to geographic data with Pelagios and Pleiades.

Tutors:

- Ryan Baumann (Duke) is a digital humanities researcher and programmer. He was a lead developer on the Son of Suda On-Line (SoSOL), Papyrological Editor, and Leiden+, to deliver scholarly editing workflow for an EpiDoc-based text corpus.
- Gabriel Bodard (King's College London) is a researcher in digital epigraphy in the Department of Digital Humanities, a member of the TEI Technical Council, and has been working on projects publishing inscriptions and papyri in EpiDoc for over ten years

(including Inscriptions of Aphrodisias, Inscriptions of Roman Tripolitania, Ancient Inscriptions of the Northern Black Sea, Papyri.info). He is one of the lead authors of the EpiDoc Guidelines and developers of the Example XSLT, and has taught regular EpiDoc training workshops in London, Rome, and elsewhere since 2005.

- Hugh Cayless (NYU) works for the Digital Library Technology Services group at NYU on projects at the intersection of ancient studies and technology. He was the lead developer on the Papyrological Navigator (papyri.info) and is currently working on standards for linked data supporting digital critical editions. He is one of the creators of EpiDoc and is a member of the TEI Technical Council.
- Henriette Roued-Cunliffe (Ludwig-Maximilians-Universität München) is a digital humanities researcher and programmer on the Buddhist Manuscripts from Gandhara project where she is using EpiDoc to create a new version of the online publication of manuscripts as well as tools for interacting with the dataset. Previously, she used a similar approach on the new Vindolanda Tablets Online II publication as a part of her PhD at University of Oxford. This involved developing the word search web service, APPELLO, which enabled the same dataset to be used in two separate applications.

Using and Customizing TEI Boilerplate

Walsh, John A.

TEI Boilerplate is an open source, lightweight and simple solution for publishing styled TEI P5 content directly in modern browsers. With TEI Boilerplate, TEI XML files can be served directly to the web without server-side processing or translation to HTML. TEI Boilerplate performs

a very simple XSLT 1.0 translation that embeds the TEI document inside an HTML shell. This embedding largely preserves the integrity of the TEI document while also allowing TEI users to use CSS and JavaScript to style the TEI content directly, manipulate TEI data, build and design interfaces, and add functionality. CSS and JavaScript skills are relatively common and widely known, and one goal of TEI Boilerplate is to provide a simple TEI publishing framework that can be used and customized by TEI users who have basic web development skills but who lack advanced XSLT knowledge. Much more detail about TEI Boilerplate—including demos, documentation, and downloads—may be found at http://teiboilerplate.org/http://teiboilerplate.org/.

The tutorial will cover basic use and configuration of TEI Boilerplate and also customization of TEI Boilerplate with CSS and JavaScript. The tutorial will include example data, and participants will also have an opportunity to work with their own data.

TEI Boilerplate was released about a year ago and remain in active development. A new 1.1 version with support for facsimile page images was just released in April, 2013. TEI Boilerplate has been adopted for TEI training, classroom use, and in a variety TEI projects.

Bibliography

- Walsh, J. A., Simpson, G., & Moaddeli, S. (2012). TEI Boilerplate. Retrieved from http://teiboilerplate.org/http://teiboilerplate.org/

Clarin, Standards and the TEI

Wynne, Martin

CLARIN is a pan-European initiative which aims to build a research infrastructure for language resources which will integrate numerous tools and resources in a distributed architecture, and which will respond to the needs of researchers across the humanities and social sciences. CLARIN

is being built on open standards, but also with a recognition that standards and guidelines are only one part of a complex jigsaw which needs to be assembled to create reliable, durable and high quality services.
A keynote speech will be given by Alexander Geyken of the Berlin-Brandenburg Academy of Sciences (BBAW) on the topic of the use of TEI in the development of the Deutsches Textarchiv.
There will be a number of presentations on topics on the appliation of the TEI guidelines to language resources and tools, and about the role of the TEI in emerging CLARIN services and standards. Presenters will not simply present an overview of their work, but focus on precisely how, why (or why not) TEI formats, guidelines and technologies are being deployed, and to go into some technical detail on these topics.
It is hoped that this will be only the start of promoting dialogue and collaboration between CLARIN and the TEI at many levels. One result would be an improved dialogue about the use of the TEI in higher-level initiatives to develop standards for the CLARIN architecture, but another would be enhanced engagement directly with the TEI community of developers and researchers in the many centres and institutions related to CLARIN.
This workshop is aimed at:

- CLARIN developers
- researchers in the humanities and social sciences already working text encoding and with CLARIN demonstrator projects
- digital humanists interested in working towards integration of their resources with the CLARIN infrastructure
- TEI members interested in developing guidelines for linguistic resources (e.g. the Linguistic SIG)

List of Authors

de la Iglesia, Martin, Göttingen State and University Library, Germany
Decorde, Matthieu, ICAR Research Lab - Lyon University and CNRS, France
Del Grosso, Angelo Mario, Istituto di Linguistica Computazionale "Antonio Zampolli" ILC-CNR, Italia
Denzer, Sandra, Technical University of Darmstadt, Germany
Driscoll, Matthew James, Københavns Universitet, DK
Dumont, Stefan, Berlin-Brandenburg Academy of Sciences and Humanities, Germany
Eide, Øyvind, University of Oslo, Norway
Embach, Michael, Stadtarchiv/Stadtbibliothek Trier, Germany; Universität Trier, Germany
Emery, Doug, University of Pennsylvania, United States of America
Fankhauser, Peter, IDS Mannheim, Germany
Fechner, Martin, Berlin-Brandenburg Academy of Sciences and Humanities, Germany
Flanders, Julia, Northeastern University, United States of America
Forsbom, Eva, Dramawebben, Sweden
Fraistat, Neil, University of Maryland, United States of America
Gavin, Michael Andrew, University of South Carolina, United States of America
Gehrke, Stefanie, Equipex Biblissima, France
Geyken, Alexander, Berlin-Brandenburg Academy of Sciences and Humanities, Deutsches Textarchiv
Glorieux, Frédéric, Université Paris-Sorbonne, France
González-Blanco García, Elena, Universidad Nacional de Educación a Distancia, Spain
Göbel, Mathias, Göttingen State and University Library, Germany
Haaf, Susanne, Berlin-Brandenburg Academy of Sciences and Humanities, Deutsches Textarchiv
Hawkins, Kevin S., University of Michigan, United States of America
Heiden, Serge, ICAR Research Lab - Lyon University and CNRS, France
Horn, Franziska, Technical University of Darmstadt, Germany
Hudrisier, Henri, Paragrpahe, Université Paris 8, France
Jolivet, Vincent, Université Paris-Sorbonne, France

Jovanović, Neven, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia
Kenny, Julia, Università di Pisa, Italy
Kossman, Perrine, Université de Bourgogne, France
Krause, Celia, Technische Universität Darmstadt, Germany
Kupietz, Marc, Institut für Deutsche Sprache, Mannheim, Germany
Lagercrantz, Marika, Dramawebben, Sweden
Lamé, Marion, ILC CNR Pisa, Italia
Lana, Maurizio, University of Piemonte Orientale, Italia
Larousse, Nicolas, TGE Adonis, France
Lavrentiev, Alexei, ICAR Research Lab - Lyon University and CNRS, France
Lemnitzer, Lothar, Berlin-Brandenburg Academy of Sciences and the Humanities, Germany
Leoni, Chiara, Università di Pisa, Italy
Lindgren, Ulrika, Dramawebben, Sweden
Magro, Diego, University of Torino, Italia
Mann, Rachel Scott, University of South Carolina, United States of America
Masotti, Raffaele, Università di Pisa, Italy
Maus, David, Herzog August Bibliothek Wolfenbüttel, Germany
Miskiewicz, Wioletta, Institut d'Histoire et de la Philosophie des Sciences et des Techniques IHPST/CNRS/Paris, Head of Archives e-LV: http://www.elv-akt.net/
Moerth, Karlheinz, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
Monella, Paolo, University of Palermo, Italy
Morlock, Emmanuelle, HiSoMA, CNRS / Université Lyon 2, France
Moulin, Claudine, Universität Trier, Germany
Muller, Charles, University of Tokyo
Muñoz, Trevor, University of Maryland, United States of America
Nagasaki, Kiyonori, International Institute for Digital Humanities / University of Tokyo, Japan
Olsson, Leif-Jöran, Språkbanken, University of Gothenburg, Sweden

Ott, Tobias, Stuttgart Media University, Germany; pagina Gmbh Tübingen
Ott, Wilhelm, Universität Tübingen, Germany
Peroni, Silvio, University of Bologna, Italia
Pfefferkorn, Oliver, IDS Mannheim, Germany
Pierazzo, Elena, Kings College London, UK
Piez, Wendell, Piez Consulting Services, United States of America
Portela, Manuel, University of Coimbra, Portugal
Porter, Dot, University of Pennsylvania, United States of America
Pugliese, Jacopo, Università di Pisa, Italy
Pytlik Zillig, Brian L., Univ of Nebraska-Lincoln, United States of America
Rapp, Andrea, Technische Universität Darmstadt, Germany
Razanajao, Vincent, Griffith Institute, University of Oxford, United Kingdom
Rindone, Francesca, Karlsruher Institut für Technology, Germany
Rodríguez, José Luis, Real Biblioteca, Madrid
Romary, Laurent, National Institute for Research in Computer Science and Control, France
Rosselli Del Turco, Roberto, Università di Torino, Italy
Roued-Cunliffe, Henriette, Ludwig-Maximilians-Universität München
Sahle, Patrick, Universität zu Köln, DE
Scacchi, Alessia, University of Rome "Sapienza", Italia
Schaßan, Torsten, Herzog August Bibliothek Wolfenbüttel, Germany
Schreiter, Solveig, Musikabteilung der Staatsbibliothek zu Berlin
Schulz, Daniela Monika, University of Cologne, Germany
Seifert, Sabine, Humboldt University Berlin, Germany
Semlak, Martina, University of Graz, Austria
Sghidi, Sihem, ISD, Université La Manouba, TUNISIE
Shimoda, Masahiro, University of Tokyo
Silva, António Rito, Technical University of Lisbon, Portugal
Steyer, Timo, Herzog August Bibliothek Wolfenbüttel, Germany
Stigler, Johannes, University of Graz, Austria
Stotzka, Rainer, Karlsruher Institut für Technology, Germany
Tomasek, Kathryn, Wheaton College, United States of America

Tomasi, Francesca, University of Bologna, Italia
Tonne, Danah, Karlsruher Institut für Technology, Germany
Trasselli, Francesca, Istituto Centrale per il Catalogo Unico, Italia
Vanscheidt, Philipp, Universität Trier, Germany; Technische Universität Darmstadt, Germany
Viglianti, Raffaele, University of Maryland, United States of America
Vitali, Fabio, University of Bologna, Italia
Walsh, John A., Indiana University, United States of America
Wiegand, Frank, Berlin-Brandenburg Academy of Sciences and Humanities
Witt, Andreas, IDS Mannheim, Germany
Wynne, Martin, University of Oxford, United Kingdom
Zghibi, Rachid, ISD, Université La Manouba, TUNISIE