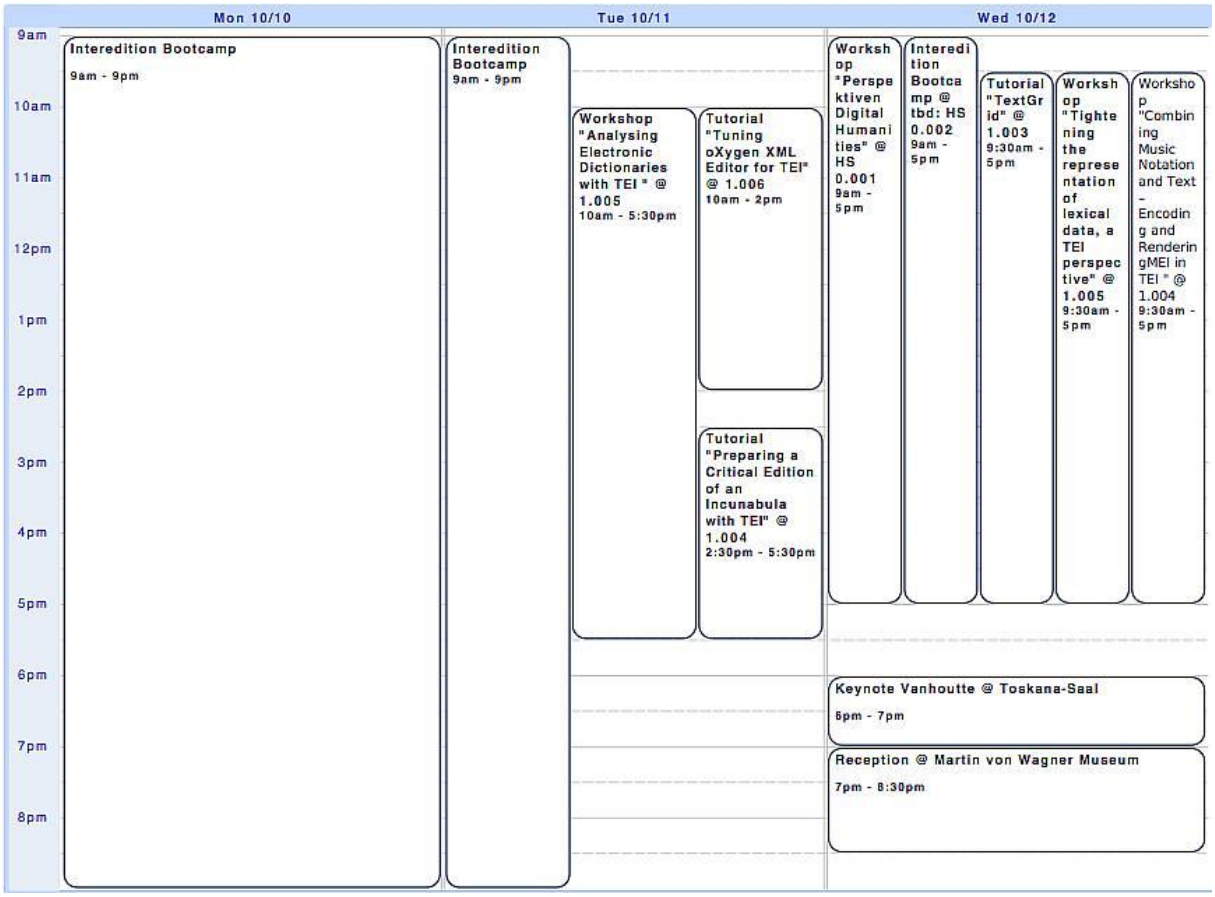
Conference Programme

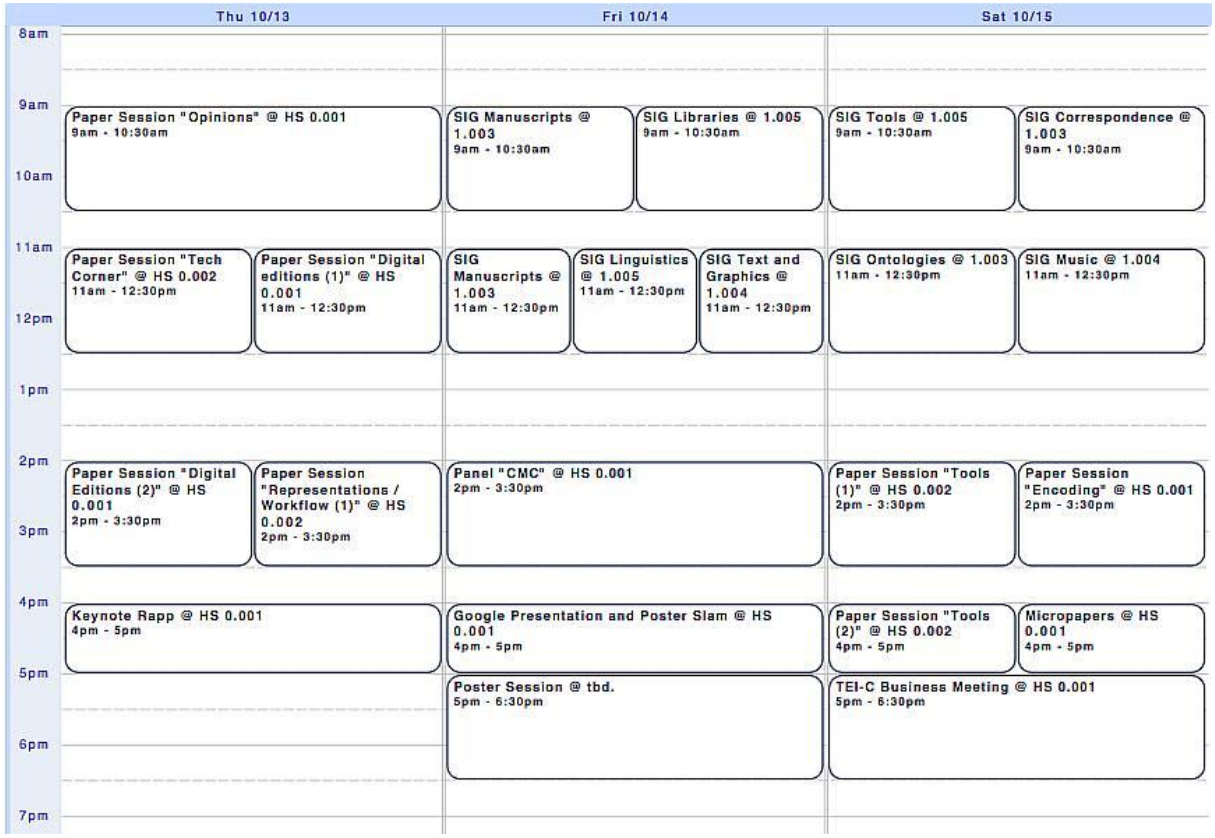
Book of Abstracts

Compiled and slightly edited by Malte Rehbein, Würzburg 4 October 2011

Table of Contents

Conference Schedule (Dates and Rooms)





Programme

Tuesday, 11 October 2011

- 10:00-17:30 Workshop "Analysing Electronic Dictionaries with TEI"
- 10:00-14:00 Tutorial "Tuning oXygen XML Editor for TEI"
- 14:30-17:30 Tutorial "Preparing a Critical Edition of an Incunabula with TEI"

Wednesday, 12 October 2011

- 09:15-16:00 Workshop “Perspektiven der Digital Humanities”
- 09:15-16:00 Interedition Think Tank
- 09:30-17:00 Workshop "Tightening the representation of lexical data, a TEI perspective"
- 09:30-17:00 Workshop "Combining Music Notation and Text – Encoding and Rendering MEI in TEI"
- 09:30-17:00 Tutorial "TextGrid"
- 18:00-19:00 Keynote Presentation "So You Think You Can Edit? The Masterchef Edition" (Edward Vanhoutte)
- 19:00-20:00 Reception (venue: Martin von Wagner Museum)

Thursday, 13 October 2011

- 09:00-10:30 Paper Session "Opinions"
- *10:30-11:00 Coffee Break*
- 11:00-12:30 Paper Session "Digital editions (1)"
- 11:00-12:30 Paper Session "Tech Corner"
- *12:30-14:00 Lunch Break*
- 14:00-15:30 Paper Session "Digital editions (2)"
- 14:00-15:30 Paper Session "Representations and Workflow"
- *15:30-16:00 Coffee Break*
- 16:00-17:00 Keynote Presentation "From text technology to cultural technology: the role of the TEI in Virtual Research Environments" (Andrea Rapp)
- Excursion: Wine tasting at the Hofkellerei

Friday, 14 October 2011

- 09:00-12:30 SIG meetings
- *12:30-14:00 Lunch Break*
- 14:00-15:30 Panel "Representing genres of computer-mediated communication in TEI"
- *15:30-16:00 Coffee Break*
- 16:00-16:30 Presentation: "Google's approach to knowledge encoding and text annotation" (Steve Crossan, Google Cultural Institute)
- 16:30-18:30 Poster Slam and Poster Session
- Excursion: Guided city tour

Saturday, 15 October 2011

- 09:00-12:30 SIG meetings
- *12:30-14:00 Lunch Break*
- 14:00-15:30 Paper Session "Encoding"
- 14:00-15:30 Paper Session "Tools (1)"
- *15:30-16:00 Coffee Break*
- 16:00-17:00 Micropapers
- 16:00-17:00 Paper Session "Tools (2)"
- 17:00-19:00 TEI-C Business Meeting

Sunday, 16 October 2011

- TEI-C Board meeting

Keynote Presentations

Edward Vanhoutte (Gent):
"So You Think You Can Edit? The Masterchef Edition"

Wednesday, 12 October 2011, 18:00h, Toskana-Saal

Culinary history shows that every wave of gastronomic innovation is directly followed by a reflective appreciation of traditional cooking techniques, terroir cuisine, and local produce. Moreover, the subtle balancing of the four basic tastes – sweet, sour, bitter and salty – and the more recently discovered fifth taste umami, together with a perfect control over texture and temperature in innovative food creations refer more to traditional cooking than newcomers in the food trade generally acknowledge. Surely, technological research and development affects the clientele's food experience because it alters the way chefs cook and dress up their dishes but it hardly replaces the achievements and insights of traditional cooking. An inclusive approach towards tradition and innovation is therefore key.

Likewise, our thinking about the digital scholarly edition should take an inclusive approach towards the accomplishments of the past. The current models for the social edition tend to underestimate the essential function of the scholarly edition in the transmission of (literary) works to next generations and focus mainly on collaboration, engagement, and participation. Claims about the 'anonymous' apparatus variorum, the failure of 'self-contained' editions and the role of the editor as 'progenitor of knowledge creation', for instance, signal some misunderstandings of traditional bibliography from the perspective of the social web. It seems that the attack on the authorial intention by sociological theories – the work as a social product – and processional theories – the work as a process rather than a product – in the 1980s and 1990s now migrates to an attack on the editorial role by the advocates of the social edition who see the scholarly text as a uniquely collaborative process defined by the available social technologies.

Therefore I wish to reiterate the importance of the 'four traditional basic tastes' of a scholarly edition – the constituted reading text, the apparatus variorum, the genetic and transmissional history, and the commentary – and add social technologies as savoriness or umami to the editorial dish. Surely, technological research and development affects the user's experience with a digital edition and it alters the ways and the conditions in which editors edit and construct their editions, but it hardly replaces the achievements and insights of traditional bibliography. An inclusive approach towards tradition and innovation is therefore key.

Biography:

Edward Vanhoutte is currently Director of Research and Publications in the Royal Academy of Dutch Language and Literature - KANTL (Gent, Belgium; and head of the Centre voor Scholarly Editing and Document Studies - CTB. He is also a Research Associate of UCL Centre for Digital Humanities (University College London). Edward is Editor-in-Chief of LLC. The Journal of Digital Scholarship in the Humanities, Managing Editor of Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde, and member of the editorial boards of Digital Studies / Le champ numérique, (SDH/SEMI) and TEI Extramural Journal-EJ. He publishes widely on (electronic) textual and genetic criticism, electronic scholarly editing, and humanities computing and he is co-author of

TEI by Example. Edward serves as a member of several boards and councils such as the Huygens Institute - ING, the executive council of the ALLC, and the technical committee of the dbnl: Digitale Bibliotheek der Nederlandse Letteren. His research interests include text-encoding and markup of modern manuscript material, electronic scholarly editing, genetic editing, and the history of electronic editing and humanities computing. Occasional blogs on Humanities Computing can be found on The Mind Tool: Edward Vanhoutte's Blog.

email: edward.vanhoutte@kantl.be | Twitter: @evanhoutte

Andrea Rapp (Darmstadt):

"From text technology to cultural technology: the role of the TEI in Virtual Research Environments"

Thursday, 13 October 2011, 16:00h, Hubland-Campus ZHSG 0.001

Virtual research environments which are supporting and at the same time altering the research process in the humanities are currently booming: they enable access to digital resources, tools and research findings in an unprecedented way. Standards, metadata and annotations are playing an eminently important role in this, and not only with regard to interoperability and sustainability of resources. For analogue representations of textual expression, cultural practices have been in the making / undergoing constant development for centuries. These practices affect the ways in which information and knowledge are stored, transmitted, explored, discovered etc. In the course of the development negotiations over (metadata-)standards, especially the TEI Guidelines, philologists have provided insights into the structure of texts, the organization of knowledge, the establishment of networks and the explicitation of this type of information. The development of bespoke cultural technologies can thus benefit from this knowlegde in order to address and establish new research questions and methods. In this paper, I am going to address and pursue these issues and challenges on the basis of concrete examples such as FuD, TextGrid, Dariah, Clarin). In particular, I will address the questions such as which cultural technologies are currently promising fruitful approaches to research questions of the humanities in the 21st century, how these can benefit from the experience of textual scholarship in the previous centuries and how current virtual research environments are likely to alter the face of textual scholarship in the future.

Biography:

Since 2010 Andrea Rapp is a professor for medieval studies and computer philology at the Department of Linguistics and Literary Studies at Technical University Darmstadt. She has been the head of the Goettingen Digitization Center at the State and University Library in Goettingen (2003-2004), afterwards one of the executive directors of the Center for Digital Humanities at Trier University (2004-2010). She has been working in the field of digitization, digital editions, electronic dictionaries, and digital humanities in general for over 20 years. She is in charge of several DFG- and BMBF-funded projects and is one of the initiators of the TextGrid project.

Website: http://www.linglit.tu-darmstadt.de/index.php?id=rapp0

Papers

Abstracts are listed alphabetically by last name of first author.

A TEI schema for the representation of CMC discourse

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika

On the basis of a comparison of several encoding options provided by the TEI-P5, this paper presents a TEI-conformant basic encoding scheme for the representation of selected genres of computer-mediated communication (CMC).

The authors of this paper are, on the one hand, corpus providers (cf. paper 2 in this panel) and, on the other hand, linguists who pursue corpus-based research on digital genres and on language use on the internet (cf. paper 1 in this panel). The encoding scheme should therefore not only meet the requirements of corpus building and integration but should also enable linguists to annotate and analyze the particular linguistic and structural properties of CMC discourse.

Thus, a TEI schema for CMC discourse that serves both purposes equally well would have to fulfill the following requirements that are addressed with a little more detail in the complete version of this abstract which can be found in the attached PDF file:

1. The schema should be adequate for rendering the specific status of CMC discourse between TEXT and CONVERSATION and take into account the written nature of the data (including the use of text design features, hyperlinks and the integration of media objects) as well as its dialogic (conversation-like) structure.
2. The schema should be compatible with the metadata and corpus data of the DWDS core corpus (cf. paper 2) which is encoded in compliance with the TEI framework and allows for a stable and persistent method of referencing the sources included.
3. It should allow for a distinction between linguistic data that has been created by users of CMC and such data that has been created by the system or a bot, as e.g., automated alerts or parts of user messages that have been automatically added during the processing of incoming user messages by the system.
4. It should allow for a separation of fragments of postings that have been originally produced by the originator from such parts that are citations of previous postings from other users. Citing other users' postings is a recurrent feature of e.g., forums/discussion boards, Tweets and e-mails.
5. In order to provide a useful basic representation as a starting point for various linguistic projects which may follow different theoretic approaches, the basic structural units should be of a kind that can easily be derived from the raw data rather than from a particular theory.
6. Nevertheless, it should include instruments for the identification and annotation of "netspeak" elements such as emoticons, acronyms, leetspeak expressions or feigned orality.
7. It should allow scholars to easily adopt, customize and extend the basic format for purposes of their individual research projects and warrant maximum interchangeability of resources which have been annotated using this basic encoding scheme.

In order to find out which encoding best meets the requirements given above, we are experimenting with several encoding options which are based on modules that are provided by the current version of the TEI guidelines: the module 4 "default text structure" (which provides a model for a broad range of text types which have been produced and edited

under monologic conditions), the module 8 “transcriptions of speech” (which provides a model for spoken dialogues) and the module 7 “performance texts” (which provides a model for written dialogues – with individual speeches assigned to several speakers – but whose concept does not comprise natural conversations).

We investigate to which extent these modules can be customized for CMC discourse. In our presentation, we will report these experiments and compare the merits and drawbacks of the different encoding options in light of the above-mentioned requirements. As a result, we will present an encoding scheme which fits well with the framework of the DWDS corpus and which makes a compromise between the requirements outlined above and the modules that are provided by the current version of the TEI.

In our examples, we will focus (1) on threads in online forums/discussion boards, on discussion pages of Wikis and on “social network” sites (asynchronous mode), and (2) on logfiles from IRC, webchat and instant messaging conversations (synchronous mode).

The encoding options as well as our suggestions for a basic TEI encoding scheme for CMC discourse will be illustrated using examples from a dataset that we retrieved from the internet in 2010 and 2011 and which includes data from a broad range of genres.

The paper will finish with an outlook on features of CMC discourse that are not yet covered by the presented format and how they could be integrated into the current annotation framework. We would like the audience to discuss the presented basic CMC encoding scheme as well as the formulated desiderata.

Bibliography:

- [TEI-P5] TEI Consortium (eds., 2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: April 24, 2011).

Representing genres of computer-mediated communication in TEI

Beißwenger, Michael; Lemnitzer, Lothar

The panel addresses issues related to a project that aims at building a TEI-compliant reference corpus of German computer-mediated communication (CMC). This corpus (‘Deutsches Referenzkorpus zur internetbasierten Kommunikation’, DeRiK) will cover a broad range of CMC genres such as e-mail, discussion boards, chats and instant messaging conversations, weblogs, wiki discussions, microblogging on Twitter and communication in “social network” sites. It shall be integrated into the DWDS corpus collection of contemporary German (Geyken 2007) and used in the context of corpus-based lexicography. In our panel we want to present data from ongoing work in our project, compare several modeling options which apply different TEI modules and discuss a draft for a TEI-based core format for the representation of CMC genres. The overall goal of the panel is to pave the way for a TEI-based encoding scheme for CMC discourse which meets the specific requirements in the DeRiK context but which may also be used by other projects that aim at building annotated corpora of CMC.

Up to now, corpus-based CMC projects have typically developed their own, project-specific encoding schemes. This complicates if not even inhibits the sharing of the data across projects. This is all the more regrettable because many projects add value to the data through their annotation. Sharing, merging and comparing datasets, particularly in constrastive research, calls for a standard-conformant basic scheme which suits the need of various projects and which is easy to handle and extend. Since many resources within the humanities use the TEI framework for annotation purposes, such a basic scheme for CMC should be conformant with TEI.

The papers of the panel will point out crucial challenges of building CMC corpora from the perspective of linguistic CMC research, discuss several options for the representation of CMC data on the basis of customized modules of the TEI-P5 and present a draft for a TEI-based encoding schema for the representation of CMC. The panel is structured as follows:

1. The first paper discusses the main structural and linguistic peculiarities of CMC discourse as described in linguistic CMC research. It will address the controversial status of CMC genres between prototypical (written, monologic) text and (spoken, dialogic) conversation and demonstrate with examples that it is thus not obvious how these specific CMC properties may best be represented in the TEI-P5 framework.
2. The second paper describes the motivation, goals and design of the DeRiK project. It focuses on the challenges that arise when integrating written CMC into an existing framework of TEI-encoded text corpora (the DWDS corpora) and outlines the requirements that result for the TEI-encoding of the CMC subcorpus.
3. The third paper compares several options for encoding CMC genres by means of (customized) modules defined in the TEI-P5 and discusses their pros and cons with respect to the challenges and requirements outlined in the papers 1 & 2. As a result, a basic TEI-conformant encoding schema for selected genres of CMC will be presented that meets both the requirements of the DWDS framework and the requirements defined from the perspective of linguistic CMC research.

The panel should be concluded by a discussion of the basic schema for CMC presented in paper 3.

Bibliography:

- Geyken, Alexander (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: Christiane Fellbaum (ed.): Collocations and Idioms. London, 23-40.
- [TEI-P5] TEI Consortium (eds., 2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: April 24, 2011).

Challenges of representing genres of computer-mediated communication in TEI: The linguistic perspective

Beißwenger, Michael; Storrer, Angelika

The paper gives an outline of the essential challenges of creating a TEI encoding scheme that captures the structure and properties of computer-mediated communication (CMC). From the perspective of linguistic CMC research and with the help of examples, we will point out that the discussion about a basic format for the representation of CMC should carefully

reflect the following issues (cf. Beißwenger & Storrer 2008) which are outlined in detail in the complete version of this abstract which can be found in the PDF attachment:

- The specific status of CMC genres between prototypical (written, monologic) text and (spoken, dialogic) conversation;
- the temporal properties of synchronous written CMC;
- the question of the basic units of the discourse structure;
- the question of CMC "macrostructures";
- the question of elements needed for the representation of linguistic features below the level of postings/turns/individual speech acts;
- the question of representing hypermedia structures;
- the question of metadata for CMC.

On the basis of outlining these issues, the paper discusses the general decisions with which one is faced when aiming at modelling CMC data using the framework provided by the TEI-P5. These general decisions are related to questions such as whether CMC discourse could be adequately represented in terms of the TEI-modules for (a) transcriptions of speech (which provides a model for spoken and not for written language), (b) text (which provides a model for a broad range of text structures which have been produced and edited under monologic conditions) or (c) performance texts (which provides a model for written dialogues – with individual speeches assigned to speakers – but whose concept does not comprise natural conversations), or whether CMC discourse, due to its crucial differences to any of the genres already recognized by the TEI-P5, should rather be treated in an own (not yet existing) module of a future version of the TEI framework.

The general issues discussed in this paper will be taken up in paper 3 and readdressed using data from several CMC genres out of the context of the DeRiK project (cf. paper 2) as well as encoding examples for these data applying several modules provided by the TEI-P5.

Bibliography:

- Beißwenger, Michael (2007): Sprachhandlungskoordination in der Chat-Kommunikation. Berlin (Linguistik – Impulse & Tendenzen 26).
- Beißwenger, Michael (2008): Situated Chat Analysis as a Window to the User's Perspective: Aspects of Temporal and Sequential Organization. In: Jannis Androutsopoulos & Michael Beißwenger (Eds.): Data and Methods in Computer-Mediated Discourse Analysis (= Special Topic Issue of Language@Internet 5). www.languageatinternet.de/articles/2008/1532/index_html/
- Beißwenger, Michael & Angelika Storrer (2008): Corpora of Computer-Mediated Communication. In: Anke Lüdeling & Merja Kytö (Eds): Corpus Linguistics. An International Handbook. Volume 1. Berlin. New York (Series: Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1), 292-308.
- Cherny, Lynn (1999): Conversation and Community. Chat in a Virtual World. Stanford (CSLI Lecture Notes 94).
- Crystal, David (2001): Language and the Internet. Cambridge.
- Garcia, Angela Cora & Jennifer Baker Jacobs (1999): The Eyes of the Beholder: Understanding the Turn-Taking System in Quasi-Synchronous Computer-Mediated Communication. In: Research on Language and Social Interaction 32(4), 337-367.

- Herring, Susan C. (1999): Interactional Coherence in CMC. In: Journal of Computer-Mediated Communication 4.4. WWW-Ressource: jcmc.indiana.edu/vol4/issue4/herring.html.
- Herring, Susan C. (Ed., 2010): Computer-Mediated Conversation, Part I. Special Issue of Language@Internet (Volume 7, 2010). www.languageatinternet.de/articles/2010
- Herring, Susan C., Lois Scheidt, Sabrina Bonus & Elijah Wright (2004): Bridging the Gap. A genre analysis of Weblogs. Paper presented at the 37th Hawaii International Conference on System Sciences. Online: doi.ieeecomputersociety.org/10.1109/HICSS.2004.1265271
- Markman, Kris (2006): Computer-Mediated Conversation: The Organiza¬tion of Talk in Chat-Based Virtual Team Meetings. Dissertation, Uni¬versity Texas at Austin.
- Murray, Denise E. (1989): When the medium determines turns: turn-taking in computer conversation. In: Hywel Coleman (Ed.): Working with Language. A Multidisciplinary onsideration of Language Use in Work Contexts. Berlin. New York (Contributions to the Sociology of Languages 52), 319-337.
- Schönfeldt, Juliane & Andrea Golato (2003): Repair in Chats: A Conversation Analytic Approach. In: Research on Language and Social Interaction 36 (3), 241-284.
- [TEI-P5] TEI Consortium (eds., 2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: April 24, 2011).
- Zitzen, Michaela & Dieter Stein (2005): Chat and conversation: a case of transmedial stability? In: Linguistics 42.5, 983-1021.

The electronic edition of the corpus written by Thomas Le Roy about the history of the Mont Saint-Michel, using the TEI

Bisson, Marie

Les curieuses recherches du Mont Saint Michel have been written by dom Thomas Le Roy and are the object of an electronic edition (work in progress). The monk wrote, for the two years he spend in the abbey of the Mont Saint-Michel (november 1646-july1647), three texts about the history of the abbey, under three different forms : an abstract of 20 pages ; a thematic texte of 200 pages and a chronological text of 600 pages. He used then of the rich library of the abbey, and follows, at least for a part, the recommendations of his congregation, the congregation of Saint-Maur.

I will demonstrate how the Text Encoding Initiative, which provides guidelines on structuring humanities and social science texts in XML (eXtensible Markup language) has been applied to the specifities of the corpus of dom Thomas Le Roy and also the project objectives. This edition should not only allow an entire version of the corpus to be consulted (until now, dom Thomas Le Roy's work had never been published in its entirety), but provides an opportunity to analyse the work of monks in the context of the Maurists historical reform in the XVIIe and XVIIIe centuries.

Although still a work in progress, the encoding of three manuscripts (BNF 13818 – twenty page manuscript written by dom Le Roy and sent to the abbey of Saint-Germain-des-Prés in July 1647 ; BNF 18950 – two hundred pages sent to the abbey of Saint-Germain-des-Prés in July 1648 ; fonds Mancel 195 – six hundred pages kept at the Abbey of the Mont Saint-Michel) allows us at this stage a glimpse of potential avenues of more precise research. This project allows a perspective on the interpretative potential that encoding presents for the corpus as it stands.

I will speak about my methodology, showing how I am using the TEI and what elements have been chosen, in regards of my hypothesis (those of the beginning and those which have appeared since). I will explain how the TEI suits my corpus in three points: normalisation, analysis and publication. I will talk about the inventory of the element I am using. I will show what tools, the XML allows me to use for the study and the analysis of my corpus (stylesheets, software, transformation tools…).

Bibliography:

- Marie Bisson, « L'édition numérique structurée des Curieuses recherches du Mont Saint-Michel de dom Thomas Le Roy », in Le patrimoine à l'ère du numérique (actes du colloque des 10 et 11 décembre 2009), C. Bougy, C. Dornier et C. Jacquemard (dir.), à paraître aux Presses universitaires de Caen.

Two sides of the same medal? Remarks on diplomatic and textual encoding in the Faust Edition

Brüning, Gerrit; Henzel, Katrin; Pravida, Dietmar

The edition of Goethe's "Faust" (https://faustedition.uni-wuerzburg.de) will provide all relevant manuscripts of "Faust" by making the facsimiles and transcriptions available. It is the aim of the edition not only to represent these manuscripts as manuscripts, but to reconstruct and visualize their genetic relations. Therefore we seek to develop new visualization strategies for the electronic medium. The user of the edition will be able to follow the genesis of one of the most important literary works in the German language in every particular detail. The edition will provide a text of "Faust" including all its drafts and an exhaustive account of the totality of the textual variances. For the encoding of the transcripts, the markup of the TEI is used. The encoding model for Genetic Editions as developed by the TEI Special Interest Group on Manuscripts (TEI MS SIG) developed an (http://www.tei-c.org/Activities/Council/Working/tcw19.html) is of substantial importance to our work.

This contribution will present how the TEI markup is used in the edition of Goethe's "Faust". Every manuscript is considered under two different perspectives: first as a piece of paper containing a material inscription and second as a medium of an abstract object, the text.

Under the first perspective the manuscript will be represented in a 'diplomatic' transcript of its inscription and in a visualization of the material structure of the bundle of papers to which a particular manuscript page either still belongs or did once belong.

Under the second perspective the representation is not necessarily limited to single manuscripts, even if the first step will always be to give a textual transcript of the inscription of single manuscripts. Furthermore, the genesis of the drama took over 60 years, and this process did not only affect the verbal shape of the work, but also its core, that is to say the conception of the work.

At the beginning, these two perspectives were intended to be combined in one complex encoding procedure within a single data file. But very soon, the fact had to be acknowledged that a combination of two widely differing perspectives leads to serious problems, especially the problem of overlapping hierarchies and/or the problem of being continuously obliged to use elements that are mutually exclusive. A special case in point is the rearrangement of segments by changing their respective positions on the manuscript.

Only the most radical solution proved to be manageable, the separation of the two levels (inscription and text) by using two separate data files. Now, both markups – the one for rendering the inscriptional record and the other one for encoding the text – can be applied without any conflict. However, the separation of the two levels on the other side does not imply that both are meant to be completely independent from each other. On the contrary, the interrelationship between both levels is of great importance for the genetic analysis.

This way of transcribing is exactly the TEI conformant model of multiple encodings of the same information described in the TEI guidelines (see chapter 20.1: www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html. As every method it has its advantages and disadvantages. The advantages, especially for dealing with Faust manuscripts, will be explained in our presentation. The splitting of the encoding rules in two different bodies of rules and concomitantly the division of most of the markup into two different types of markup shall be summarized. It will be illustrated with examples that will give some more insight into the asserted necessity of distinguishing the two perspectives.

And finally there are many questions of how to deal with the obvious disadvantages that come along with the division of transcripts, that is to say the "the maintenance of multiple copies of identical textual content" as well as the the missing "explicit indication that the various views, which might be in separate files, are related to each other" (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHME). What are the practical consequences of separating the diplomatic from the textual transcript for further steps in generating the edition and how will it be possible to avoid inconsistency? How are we going to evaluate and how to relate all information distributed on both levels? What do we have to keep in mind for the implementation of the genetic reconstruction?

Bibliography:

- https://faustedition.uni-wuerzburg.de
- http://www.tei-c.org/Activities/Council/Working/tcw19.html
- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html

Improving the Usability of Corpus Markup and Analysis Tools by Studying their Presentation Layer

Burghardt, Manuel; Fuchs, Markus; Wolff, Christian

The TEI plays an outstanding role as a first approach towards a standardized representation of annotations. While the standardization of the representation (encoding) of annotations has steadily evolved over the last years, the presentation or visualization of markup and its

implications for the usability of markup tools has been treated with significantly less attention. In many cases the representation of markup and its actual presentation to the user are but the same thing: plain text (original text) and markup tags (annotation) at the code level. Peter Flynn has observed that “markup experts” have a different idea about the structure of a text than “conventional writers”, the one group seeing a document as a hierarchical tree with different kinds of nodes, and the other group seeing a text as a “continuous linear narrative, broken into successive divisions” (Flynn 2009). This divide between markup experts and plain annotators is of particular importance for the case of the TEI, which was designed as a representation standard for the humanities, social sciences and linguistics. Due to the research tradition and prevailing methods in this field, humanists often lack deep technical skills, i.e. many of them aren’t aware of basic markup concepts such as document types or document trees. At the same time, it has become clear that tool and ICT usage is as indispensable for the humanities as for any other field of research (Toms & O’Brien 2008). Santos & Frankenberg-Garcia claim that “most existing corpora today are only available to and understood by a small, restricted community of users” (Santos & Frankenberg-Garcia 2007). This makes usability and user experience on the presentation side a vital component for markup- and analysis-tools. Consequently, several tools try to hide the actual representation of markup from the user, by providing different interface designs and visualizations of the data. Unfortunately, current approaches to these issues are often not in accordance with existing usability standards like ISO 9241-110:2006 for dialog principles or ISO 9241-151:2008 for the usability of web interfaces (Dipper et al. 2004, Burghardt & Wolff 2009).

We identify two major challenges for the presentation layer of markup software which should be considered by tool designers in order to enhance the acceptance and the actual usage of standardized markup like the TEI guidelines, and to prevent corpora from becoming expensive data graveyards (Soehn et al. 2008), as corpus creation and especially intellectual annotation are extremely cost- and labor-intensive tasks. These challenges are at the same time general and domain-independent requirements for tools which strive for a high level of usability and user experience. The first requirement is an adequate visualization of data and annotation, the second requirement calls for appropriate interface and interaction design for markup- and analysis-tools. These requirements affect different stages in the typical workflow for the creation and use of corpora, which we call corpus pipeline. The corpus pipeline describes all steps necessary to fulfill an information need by querying an annotated corpus, starting from the creation and annotation of the actual corpus and ending with the query building and visualization of results. As presentation and representation often can’t be separated precisely during the first two stages (“digitization” and “normalization”) the presentation of markup is mainly an issue in the succeeding stages: “annotation”, “query building” and “visualization of results”. In the paper, we will derive and explain specific requirements for the presentation and interaction layer of each of these three stages (e.g. visualization of original text and multiple layers of annotation as well as the underlying annotation scheme during the annotation stage) by looking at existing tool solutions and by comparing different user interface design models with each other (recent examples of visually enriched tools including e.g. WordTree (Wattenberg & Viegas 2008) and DoubleTree (Culy & Lyding 2010) in the stage of “visualization of concordances”).

We argue for a user-centered presentation of markup, starting from the annotation of text, and ending with the querying of a corpus of documents and the presentation of query results. Future work will include a detailed user study and evaluation of different presentation aspects, such as the best presentation of multilayer annotation or complex queries.

Bibliography:

- Fuchs, Markus (2010). Aufbau eines linguistischen Korpus aus den Daten der englischen Wikipedia. 2010. In: Pinkal, M. & Rehbein, I. & Schulte im Walde, S. & Storrer, A. (Hrsg.).
- Semantic Approaches in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2010 (KONVENS 10). Saarbrücken: Universitätsverlag des Saarlandes. S. 135-139. Online verfügbar unter: <http://wikicorpus.com/ErstellungWPKorpus.pdf>
- Burghardt, Manuel & Christian Wolff (2009). Werkzeuge zur Annotation diachroner Korpora. In: Hoeppner, Wolfgang (Hrsg.).
- Proc. GSCL-Symposium Sprachtechnologie und eHumanities. Technische Berichte der Abteilung für Informatik und Angewandte Kognitionswissenschaft, 2009-01. Abteilung für Informatik und Angewandte Kognitionswissenschaft, Universität Duisburg-Essen, Duisburg, S. 21-31.Online verfügbar unter: <http://epub.uni-regensburg.de/6756/>
- Burghardt, Manuel & Christian Wolff (2009). Stand off-Annotation für Textdokumente: Vom Konzept zur Implementierung (zur Standardisierung?). In: Chiarcos, Christian et al. (Hrsg.).
- Von der Form zur Bedeutug: Texte automatisch verarbeiten. Proceedings of the Biennial GSCL-Conference 2009 in Potsdam, S. 53-59. Online verfügbar unter: <epub.uni-regensburg.de/14223/>

Glossing music theory: how to make transparent the web of quotations, authorities and allusions in medieval texts

Desmond, Karen

This paper takes as its starting point a music theory text known as the *Ars nova*. This text has been considered foundational to our understanding of medieval music history. In the fourteenth century there was a profound shift in musical style from the previous century's *ars antiqua* (the "old art") to what was termed the *ars nova* (the "new art"). The *Ars nova* was the medieval "avant-garde" with a sound that combined new rhythms, harmonies and texts in complex structural and formal layers. This complexity was due in large part to the expansion and reformulation of the musical notation system. The *Ars nova* theory treatise was a short technical manual that contained rubrics on how to interpret this new notation system. In the traditional historical narrative, the supposed author of this treatise was the composer and poet, Philippe de Vitry (1291-1361), who wrote music in the new style, and was quickly crowned through the annals of music history as the figurehead and putative creator of the ars nova movement.

This narrative is extremely simplified. There is no "one" complete text of the *Ars nova*, but in fact, a small handful of related, but widely divergent, texts extant in manuscripts dating from the fourteenth and fifteenth centuries. The sixteen or so texts that present these new notational theories vary in many ways: in levels of completeness (many of the texts start or

break off mid-treatise), in the order of topics presented, in prose style (for example: discursive vs. bulleted-list), and in the textual content itself, that is, the actual words and phrases used to describe specific concepts. There is in fact little hard evidence that proves that Philippe de Vitry actually penned a treatise called the *Ars nova*, and the extant texts may in fact represent remnants of a fluid teaching tradition that may or may not have originated with Vitry. Editions of many of these texts may be found today in various edited volumes, journal articles dating from 1908, 1929 and 1958, and in the nineteenth-century collection *Scriptores de musica medii aevi* edited by Edmond de Coussemaker. However, the various presentation formats, specific editorial policies and accessibility issues have obfuscated attempts at the analysis and interpretation of these texts.

In this paper, I discuss how the complex web of relationships between these sixteen texts may find its best representation in digital form. I focus my discussion on a digital edition I am preparing of three texts. following TEI guidelines, from the *Ars nova* tradition found in these manuscript sources (US-Cn 54.1; E-Sc 5.5.25; I-Su L.V.30). I plan to include annotations that link these texts to the thirteen other related texts of the *Ars nova*, following the Open Annotation Collaboration Data Model. This edition will also have an impact on the broader field of medieval studies as it offers a small-scale study of how the web of quotations, authorities and allusions in medieval texts could be made more transparent and accessible through the use of these types of digitals tools, such as the annotation model. The current process of discovering relationships between texts relies to a large extent on serendipitous discoveries in the footnotes of scholarly articles. The rate of discovery and the level of analysis of the medieval web of textual allusion would increase exponentially with the increased availability of electronic editions, especially if these editions are overlaid with annotations recording and mapping out relationships between texts. It is hoped that this paper will offer an example of how to present this particular textual tradition and others like it.

Bibliography:

- Karen Desmond, "'Secundum istos quorum nunc doctrina sequimur': The Tonary of Jacobus de Montibus," in Mosan Voices: Musical Practices, Communities, and Diasporas of the Liège Diocese (12th -18th century), ed. Catherine Saucier and Pieter Mannaerts (forthcoming, 2011).
- Karen Desmond, "Behind the Mirror: Revealing the Contexts of Jacobus's Speculum Musicae," Ph.D. diss., New York University, 2009.
- Karen Desmond, "New Light on Jacobus, Author of Speculum musicae," Journal of Plainsong and Medieval Music 9 (2000), pp. 19-40.
- Karen Desmond, "Sicut in grammatica: Analogical Discourse in Chapter 15 of Guido's Micrologus," Journal of Musicology 16 (1998), pp. 467-493.

Solving Problems for Online Diplomatic Editions of Medieval Manuscripts

Fredell, Joel Willis; Borchers, Charles W.; Ilgen, Terri Jo

Many special characters in medieval manuscripts have created problems for online transcription, including glyphs for contractions, medieval punctuation, bracketing, and other non-alphabetical elements. Unicode characters for these elements exist, if at all, only in

private forms such as MUFI. Even relatively common characters such as thorn and yogh can easily turn into empty boxes depending on the browser. Up to now, consequently, scholars hoping to publish more than a simplified and normalized transcription of a manuscript in digital form have relied on CDs, on which they can load custom fonts. Even this strategy, though, does not offer the choice of viewing a transcription with contractions (a pervasive feature in many medieval manuscripts) in the original glyphs or expanded. Furthermore, coding XML documents in TEI for such a choice is quite onerous in manuscripts where scribes use many contractions—as much as doubling the coding for a new online facsimile and transcription of British Library MS Add. 61283, the sole witness to The Book of Margery Kempe. The coding team for this project have uncovered strategies that solve both those problems: automating the XML workflow, and embedding an open-source font.

Solution 1: XML Automation

Among the problems for coding contractions: some occur with high frequency, others are occasional at most. The team developed a method to automate a first pass for coding: 1) using Stéfan Sinclair and Geoffrey Rockwell's Voyeur Tools, the team identifies which words occur with the most frequency in the manuscript; 2) the team compiles, from Voyeur Tools analysis, a custom find-and-replace-with-code catalog; and 3) the team imports this catalog into DigitalVolcano's freeware TextCrawler, which is then used to transform the initial transcription into Oxygen-ready coded form—automating the team's "first pass" and encoding in seconds what would have taken the team weeks, if not months, to do without the software.

Solution 2: Medieval Fonts

Encoding medieval manuscripts for diplomatic transcription in XML is challenged by scribes' use of certain characters (e.g. the thorn [ ], yogh [ ], punctus elevatus [ ], punctus elevatus diagonalis [ ], punctus versus [ ]) for which there may be neither Unicode character nor actual font support.

For scholarly works created for publication in print and/or for distribution via common document types (e.g. DOC, PDF, RTF, WPD) and/or electronic media (e.g. CD, DVD), this problem has largely been resolved by the Medieval Unicode Font Initiative (MUFI)'s character recommendations—the latest of which, Version 3.0, proposes the appropriation or addition of 1,548 characters from/to Unicode for use by medievalists—and through the cooperation of font developers, who have created fonts (e.g. Junicode, Andron Scriptor Web) both supporting MUFI's character recommendations and embeddable within these documents and/or distributable on (and, thereby, accessible from) these electronic media.

For scholarly works created for publication to the Web, however, the problem remains largely unresolved. Medievalists have had either to provide as a download to their sites' visitors the font(s) supporting the medieval characters used in their scholarly work or to find a way of representing these medieval characters graphically (e.g. as GIFs, JPGs, PNGs, SVGs), as opposed to typographically. Downloads require documentation and technical support when they fail to result in the successful installation of the font(s). And graphical representation can be both tedious (e.g. for each character, a second, third, or fourth

graphical copy may be required to represent its later addition or deletion or appearance in a different size later in the scholarly work) and require certain compromises on the part of the medievalist (e.g. in terms of how these graphical characters should appear when/if the Web page on which they appear is zoomed and/or when only the text on that page is zoomed).

In wrestling with the problem while encoding The Book of Margery Kempe for scholarly publication to the Web, Southeastern's Kempe Project Team devised a method for embedding MUFI-compatible fonts directly within their project's Web site.

Their method relies upon an understanding of 1) the process(es) through which different Web browsers (e.g. Internet Explorer, Safari, Firefox, Chrome, Opera) can render fonts in a Web page; 2) the types of fonts (e.g. EOT, OTF, SVG, TTF) that can be rendered by these different Web browsers; and 3) how widely different even fonts identified by the same name may be from application to application (e.g. Corel WordPerfect to Microsoft Word to OpenOffice.org Writer) and operating system to operating system (e.g. Linux to Mac to Windows).

Bibliography:

- Margery Kempe, The Book of Margery Kempe: A Manuscript Facsimile and Diplomatic Edition, in collaboration with the British Library. Online edition. [forthcoming]
- “Alchemical Lydgate.” Studies in Philology, 107 (2010): 429-64.
- “The Gower Manuscripts: Some Inconvenient Truths,” Viator 41 (2010): 231-50.
- “Design and Authorship in the Book of Margery Kempe.” Journal of the Early Book Society, 12 (2009): 1-34.

Metadata customization with ODD

Gaiffe, Bertrand François

In the Clarin project[3], a flexible metadata scheme has been proposed [2, 1]. MPI proposed an implementation 1 that relies on W3C schemas, but according to us, this implementation lacks flexibility. We thus proposed an alternative ODD-based implementation. Unfortunately, ODD lacks two essential features for this task: cardinalities restrictions and interleaving. In this abstract, we will describe the initial problem and sketch the two extensions we added to ODD.

Bibliography:

- D. Broeder, T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari, and P. Wittenburg. Foundation of a component-based flexible registry for language resources and technology. In N. Calzorali, editor, Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pages 1433-1436. European Language Resources Association (ELRA), 2008.
- D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In J. Mariani J. Odjik K. –Choukri, S. Piperidis M. Rosner N. Calzolari, B. Maegaard and D. Tapias,

editors, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pages 43-47. European Language Resources Association (ELRA), 2010.
- T. Vradi, P. Wittenburg, S. Krauwer, M. Wynne, and K. Koskenniemi. Clarin: Common language resources and technology infrastructure. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) 2008.

A large scale critical edition: first translation od St Augustine's City of God by Raoul de Presle

Gaiffe, Bertrand François; Stumpf, Béatrice

An important part of the vocabulary of politics in French language comes from medieval translations from Latin and Greek into French during the 14th and 15th centuries. This period favoured neologisms in the political science field because translators faced Latine or Greek concepts which did not exist yet in French [1]. This corpus of medival translations is only partially edited. In particular, City of God, written by Augustine of Hippo and translated and commented by Raoul de Presles, which is a major work in the history of ideas in the West, was never edited until now. In order to study the History of the French Political Science Lexicon (HFPSL, acronym of the project by Erc), our team took in charge the edition of this huge text. The main manuscript belonged to the royal library of Charles V, (BnF: Biblioth`que nationale de France, department of Western Manuscripts, Fr 22912 e (P1) and Fr 22913 (P2)).

The edition of the 894 folios of our main manuscript will allow researchers to have an access to this text in order to lead new researches in linguistics, history and political sciences, and more generally in Humanities. In the paper, we describe the characteristics of the edition, the TEI encoding and the tools developed.

Bibliography:

- Olivier Bertrand. Le vocabulaire politique aux 14e et 15e si`cles: constitution d'un lexique ou mergence d'une science ? In Olivier Bertrand, Hiltrud Gerner and Béatrice Stumpf, editors, Lexiques scientifiques et techniques.Constitution et approches historiques, pages 923. Editions de l'Ecole poly-technique, Palaiseau, 2007.
- Groupe de recherches. " La civilisation de l'écrit au Moyen Age ". Conseil pour l'édition des textes médiévaux, fasc. 3, Textes littéraires. Comité des travaux historiques et scientifiques, Ecole nationale des chartes, 2002.
- Groupe de recherches. " La civilisation de l'écrit au Moyen Age ". Conseil pour l'édition des textes médi´vaux, fasc. 1, Conseils généraux. Comité des travaux historiques et scientifiques, Ecole nationale des chartes, 2005

The Canary in the Text Mine: Analysis of the data mining of TEI-encoded texts in MONK research software

Green, Harriett Elizabeth

TEI is one of the most developed tools for analyzing texts on the micro-level and for data mining a large mass of texts. Yet how is TEI-enhanced software being utilized by humanities

scholars for their research? This paper presents analysis on the use of MONK, a text-mining software that utilizes TEI-encoded texts to facilitate quantitative analysis of literary texts. The study examines the research conducted in the database using twelve months of website transaction logs from 2010 and series of interviews with researchers who use MONK in their research.

BACKGROUND

MONK is a text mining research tool hosted by the University of Illinois at Urbana-Champaign Library that enables humanities scholars to mine data from TEI-A encoded texts in select literary databases and archives. MONK builds upon two previously developed text mining programs NORA (http://www.noraproject.org/) and WordHoard (http://wordhoard.northwestern.edu/) in order to create a powerful new environment that "lets users carry out complex data-mining and query operations across collections that contain nearly 200 million words" (MONK documentation, www.monkproject.org/background.html). The SEASR (http://seasr.org/) environment provides the tools for statistical analyses in MONK.

MONK contains a selection of texts spanning from the sixteenth century through late nineteenth century that are encoded in TEI-Analytics or TEI-A, a TEI markup specially created for analytics, via the Abbot tool. Abbot ingested the TEI source files for the texts and normalized them into TEI-A Researchers can also encode and import texts into MONK with the use of Zotero and a MONK Firefox extension.

DATA AND METHODOLOGY

Statistics about the web traffic and usage of MONK were gathered using AWStats, a web log analyzer used to track the web statistics for MONK. The statistics used in this study were gathered January 2010 through December 2010. The investigator is also conducting interviews and surveys with researchers who use MONK, and the analysis of the qualitative data will be completed this summer.

The statistics analyzed include the number of visits on each webpage within MONK, the amount of data processed through each page, number of entry and exit visit, length of the visits, and users' geographic locations.

The statistical analyses conducted on the data included calculating the mean of users that accessed each section of MONK; the mean amount of data transmitted through each page; the distribution of accessed MONK webpages, which were coded as Orientation, Workset, and Toolset webpages; and the frequency of entry and exit points among these three types of webpages.

ANALYSIS

The quantitative analysis has revealed that the text mining tools in MONK are primarily being used to compile work sets and conduct preliminary statistical analysis. The most frequent tools used on average were:

- https://monk.library.illinois.edu/secure/get/CorpusManager.getWorkList = compiling the worksets
- https://monk.library.illinois.edu/secure/get/ProjectManager.getToolSets = selecting a toolset
- https://monk.library.illinois.edu/cic/public = the opening page

One hypothesis that might be drawn from this data points is that many users are in the initial exploratory steps of using MONK by creating accounts and putting together their first worksets to analyze. Another point of note is a comparison of the accessed MONK pages, which reveals that the use of analytics toolsets and the use of tools for creating worksets were accessed by researchers at a proportion of 2 to 1. In another analysis, the largest amount of data was utilized for a tool comparing the frequency of word features, with a usage of 272.1 MB on average and 15% of the total data processed. These are only a sample of the analyses conducted so far.

This initial examination has begun to reveal several early insights on how scholars are conducting textual analysis research in MONK, and how TEI-A is a critical component of conducting data mining across mass texts. Ultimately, this study critically reveals new avenues of analyzing the research use of TEI in generating quantitative data for textual analysis, and how TEI can be leveraged even further in digital humanities tools.

Bibliography:

- American Council of Learned Societies. (2006). Our Cultural Commonwealth: the Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. New York: author.
- Friedlander, A. (2009). Asking Questions and Building a Research Agenda for Digital Scholarship. In Council for Library and Information Resources, Working Together or Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and the National Endowment for the Humanities. Washington, D.C.: Author.
- Khanal, N., Kehoe, A., Kumar, A., MacDonald, A., Mueller, M., Plaisant, C., Ruecker, S., Sinclair, S. & Unsworth, J. (2009). MONK Tutorials. Retrieved from monkpublic.library.illinois.edu/monkmiddleware/public/index.html
- Pytlik-Zillig, B. L. (2009). TEI Analytics: converting documents into TEI format for cross-collection text analysis. Literary and Linguistic Computing. 24, 187-192.
- Sinclair, S. (2003). Computer-assisted reading: Reconceiving textual analysis. Literary and Linguistic Computing. 18, 175-184.
- Sperberg-McQueen, C. M. (1991). Text in the electronic age: textual study and text encoding, with examples from medieval texts. Literary and Linguistic Computing. 6, 263-279.
- Warwick, C. (2004). Print Scholarship and Digital Resources. In S. Schreibman, R. Siemens, & J.Unsworth (Eds.), A Companion to Digital Humanities, Oxford: Blackwell.

Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text

Haaf, Susanne; Geyken, Alexander

This paper presents an extensive and complex approach for the analysis and correction of double-keying errors, which is currently applied by the DFG-funded project "Deutsches Textarchiv" in order to evaluate and increase the correctness of text transcriptions and annotations of historical text. Statistical analyses of the error detection and correction results based on a large amount of analyzed text will be presented in order to verify and specify the common accuracy rates for the double-keying method.

CMC as a component of a balanced, TEI-encoded corpus representing contemporary German: goals, motivation, design issues

Lemnitzer, Lothar; Geyken, Alexander; Beißwenger, Michael; Storrer, Angelika

In this paper, we will present DeRiK ('Deutsches Referenzkorpus zur internetbasierten Kommunikation'), a common initiative at TU Dortmund University (Michael Beißwenger and Angelika Storrer) and the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW; Alexander Geyken and Lothar Lemnitzer). The goal is to produce a module of sufficient size and diversity of Computer-Mediated Communication (CMC) as a complement to the reference corpus of the DWDS project.

Since all resources of the DWDS are encoded in compliance with the TEI standard, we want to use and customize TEI for the appropriate base-level annotation of the CMC sub-corpus, thus allowing users of all types to supply their own information on top of this base encoding.

In our paper, we will address the challenges and problems of integrating the CMC component into the DWDS framework and discuss the potentials and the restrictions of the encoding options provided by the TEI-P5. It is our goal to define an encoding scheme for CMC genres which serves the lexicographical requirements that arise from the work at BBAW as well as the needs of linguistic CMC research that have been outlined in paper 1 in this panel. Suggestions for solving some of the problems discussed here will be presented in paper 3 of the panel.

Bibliography:

- Geyken, Alexander (2005): Das Wortinformationssystem des Digitalen Wörterbuchsder deutschen Sprache des 20. Jahrhunderts (DWDS). In: BBAW Circular 32. Berlin.
- Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Christiane Fellbaum (ed.): Collocations and Idioms. London, 23-40.
- Geyken, Alexander & Thomas Hanneforth (2006): TAGH – A Complete Morphology for German based on Weighted Finite State Automata. In: Proceedings of FSMNLP 2005, 55-66.
- Jurish, Bryan (2003): A Hybrid Approach to Part-of-Speech Tagging, Final report, Project 'Kollokationen im Wörterbuch', Berlin-Brandenburgische Akademie der Wissenschaften, Berlin.

- Ooi, Vincent (2002): Aspects of computer-mediated communication for research in corpus linguistics, in: Pam Peters, Peter Collins & Adam Smith (eds.): New Frontiers of Corpus Research. Amsterdam. New York, 91-104.
- Sokirko, Alexey (2003): DDC – A search engine for linguistically annotated corpora. In: Proceedings of Dialogue 2003, Protvino, Russia, June 2003.
- [TEI-P5] TEI Consortium (eds., 2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: April 24, 2011).
- van Eimeren, Birgit & Beate Frees (2010): Fast 50 Millionen Deutsche online – Multimedia für alle? Ergebnisse der ARD/ZDF-Onlinestudie 2010. In: Media Perspektiven 7-8(2010), 334-349.

Collaborative & non-deterministic markup: the CLÉA project

Meister, Jan Christoph; Petris, Marco

Markup seems on the downturn—the more comprehensive our digital collections of humanistic artefacts become, the higher the success rate of automated analysis. Instead of pre-categorizing and tagging texts in terms of human defined high-level criteria and taxonomies we are now able to retrieve relevant results from the raw data hic et nunc with the help of a tokenizer and some mathematical wizardry. In many users' everyday digital practice authoritative, rigid top-down directories have long been replaced by the extremely flexible, multi-variable bottom-up oracles of Google et al. which no longer force us to conceptualize our field of interest along somebody else's lines—or so it seems. Is this where we're about to go in DH, and in particular in literary computing? Is TEI just an encyclopaedic rear guard action trying to fight off the stochastic forces? As far as expository (non-aesthetic and non-fictional, domain specific) texts go: may be. As far as literary texts are concerned: certainly not. In this paper, we will

- try to make a case for collaborative markup
- a non-deterministic approach to markup within the TEI framework
- demonstrate the underlying data model and concept of hermeneutic markup as implemented in the current CLÉA-project.

Bibliography:

- 'Computerphilologie.' In: Gerhard Lauer, Christine Ruhrberg (eds.): Lexikon Literaturwissenschaft. - Hundert Grundbegriffe. Stuttgart (Reclam) 2011, 54-56.
- Stefan Gradmann, Jan Christoph Meister: "Digital document and interpretation: re-thinking 'text' and scholarship in electronic settings". In: Poiesis & Praxis. International Journal of Ethics of Science and Technology Assessment, 2008. Electronic pre-publication: www.springerlink.com/content/g370807768tx2027/fulltext.html
- Crowd sourcing "true meaning": A collaborative markup approach to textual interpretation. In: Harold Short, Marylin Deegan (eds.): Collaborative Research in the Digital Humanities. Festschrift for Harold Short. Ashgate Publishing Ltd., Surrey: 14 pages; in print (2011)

Faust: Multiple Encodings and Diplomatic Transcript Layout

Middell, Gregor; Wissenbach, Moritz

The central concern of the edition of Goethe's Faust (https://faustedition.uni-wuerzburg.de) is the exposition of the work's genesis. In accordance with established editorial practice, a distinction between different levels of degree of interpretation is made: The presentation of the ''record'' enables a reader to follow and judge the editorial ''interpretation''. In our case, the representation of the record comprises a detailed account of material aspects of a manuscript and the topography of its inscription. The representation of the editorial interpretation comprises encoding of textual structure, textual modifications and inter-document genetic relations.

In the process of encoding, it quickly became evident that the structures of the two views on the text are disparate, which is to say in terms of markup, they overlap or do not nest properly. This is a well-known problem in text encoding. It is discussed and several solutions are suggested in the TEI Guidelines (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html). Employing a workaround technique such as using ''milestone elements'' would have sacrificed many of the benefits of XML, among the most important of which would have been formal validation and human readablility and processability. For this reason, we decided to encode both levels as seperate XML files, which are to be combined during processing automatically. This approach is mentioned as ''Multiple Encodings'' in the Guidelines (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHME). The challenges are to develop a suitable collation algorithm and implementation which correlates parts of the two files, as well as a suitable intermediate data structure that holds the results and can be queried adequately.

The presentation of the manuscripts provides a basis for the genetic analysis and as such must be carried out with care and in a detailled fashion. First, a high-resolution facsimile of the manuscript will be made accessible. Second, a diplomatic transcript lowers the barrier to reading the 18th- and 19th-century manuscripts. On this basis, a more elaborate editorial reconstruction of the genetic process is provided.

The encoding of the manuscript transcripts follows the TEI Guidelines in combination with a recent proposal of the Special Interest Group on Manuscripts (http://www.tei-c.org/Activities/Council/Working/tcw19.html). It captures aspects of the textual structure as well as the materiality and topographical layout of the manuscript. From this information, a diplomatic transcript is automatically produced. It presents the manuscript not in its immediate material and topographical conditions, but in a deliberately idealised form. The talk will discuss challenges and solutions to encoding and presenting manuscripts with features such as erasures, revision groups, multiple writers, transpositions of written text, graphical marks, different writing directions and overlapping inscription. The first challenge presented by this approach is the encoding of complicated manuscripts in a standardised way; the second challenge is the evaluation of layout constraints to produce a pleasing diplomatic transcript.

Bibliography:

- https://faustedition.uni-wuerzburg.de
- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html
- http://www.tei-c.org/Activities/Council/Working/tcw19.html

Creating lexical resources in TEI P5. Experiences from building multi-purpose digital dictionaries

Moerth, Karlheinz; Budin, Gerhard

While using the TEI dictionary module to encode digitized print dictionaries has become a fairly uncontested and very common standard procedure, using the very same system for NLP purposes is quite another story. Our paper will report on a project creating glossaries and dictionaries which are intended to be usable both for human readers and particular NLP applications. It will comprise two parts: in the first section, the authors will try to answer the question why they use the TEI dictionary module as their preferred means to go about the task, also discussing standards and tools such ISO TC 37's Data Category Registry (ISOCat).

The second part will attempt to pinpoint some encoding issues arising from the projects under discussion such as e.g. keeping track of production metadata or how to include corpus derived statistical data. Another important issue is internal linking and how to reuse examples at various points in the dictionary. We will try to show how the produced data can be delivered as part of a service oriented lexical information system.

In the world of digital dictionaries a great number of different formats coexist: MULTILEX, GENELEX, OLIF, MILE, LIFT, OWL, ISO 1951, LMF (ISO 24613:2008) and the TEI's chapter on dictionaries whose authors had a very wide range of different applications in mind:

"... The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for wordlists, lexica, glossaries, etc. included within other documents." (TEI P5 p.251)

The ICLTT's collection of digital textual resources also comprises some smaller dictionaries/glossaries which are mainly of historical interest. Currently, efforts are being made to make this data P5 compliant. However, apart from digitizing paper dictionaries, the department has also started a second line of development creating digital lexical resources which–in part–build on the department's large digital text collections. A number of monolingual and bilingual glossaries and dictionaries are being prepared which are being compiled for particular specialised purposes. Among these resources, there count a glossary of Austriacisms (by which we understand words or phrases considered typical of the German variety spoken in Austria), and a comprehensive dictionary of modern Persian single word verbs. Both resources are being worked on at the moment and both are designed to be used in two scenarios: (a) serving as source to be queried and read by human users in a web-interface and (b) providing data that can be used in NLP applications. The Glossary of Austriacisms is planned to be utilized by tools performing automatic lexical analyses of the

Austrian Academy Corpus, a large digital text collection being maintained by the ICLTT. The above mentioned specialised verb dictionary is supposed to furnish data for the creation of a full-form lexicon which in turn is intended to be applied in a morphological analyzer.

Bibliography:

- Karlheinz Moerth, Gerhard Budin, Heinrich Kabas: Towards finer granularity in metadata: Analysing the contents of retro-digitised periodicals (Under review for jTEI)
- Wolfgang U. Dressler, Karlheinz Mörth: Produktive und weniger produktive Komposition in ihrer Rolle im Text an Hand der Beziehungen zwischen Titel und Text. (Forthcoming in "Linguistik - Impulse und Tendenzen" (De Gruyter))
- Recent conference papers: Karlheinz Mörth, Niku Dorostkar, Alexander Preisinger: Gleaning micro-corpora from the internet: integrating heterogeneous data into existing corpus infrastructures. Presented at CILC (III Congreso Internacional de Lingüística de Corpus, Valencia) 2011.
- Karlheinz Moerth, Matej Durco: In quest of a multi-purpose multi-corpus service based corpus research tool. Presented at PALC (Practical Applications in Language and Computers, Lodz) 2011.

The TEI encoding of textual fragments : dangerous wager or efficient stratagem ?

Morlock-Gerstenkorn, Emmanuelle

This paper attempts to question the encoding of textual fragments using the TEI guidelines. The fragment corresponds to part or a (small) portion of a whole that is missing. Whether it is a rest of an object that disappeared or an unfinished embryo of a work in progress, the fragment is transmitted to us disconnected from the complete and finished opus that would give him his nature, function, and finality. Since the TEI is an encoding scheme that views the text as an ordered hierarchy of content objets (OHCO), as it has been analyzed thorougly, it's not possible to use it without giving each element a tag situated in that hierarchy, and therefore its semantics and functionality. In that perspective, one can ask wether or not the choice of the TEI as en encoding scheme can be misleading and produce as a result improper interpretations ?

But above all, editing fragments consists in establishing it in a set that will determine the way they are read and interpreted. A new signification will be necessarily induced by the new configuration. Can this presentation bias that promotes groundlessly one order in a textual hierarchy superior to the others be avoided? The solution may be found in the dynamic edition, the one that can offer every possible presentation without imposing one as more important than the others. The critical electronic edition of the documentary files of Gustave Flaubert's last novel"Bouvard et Pécuchet" relies on that viewpoint. This project aims to propose an edition that could give the fragments of citations he collected and started to organize the mobility they deserve, as the volume was very far from finished when he died.

For this project, the TEI is used very pragmatically with two goals. The first consist in "recording" a base structure corresponding to the way these fragments are scripted on the

pages of the manuscript. The second is to use it as a base for en extraction of the editorial units the edition will present away from the original context of the page.

This strategy was only possible because the abstract models of the inscription of the fragments and of the edition that has to be made were clearly established. It shows that strongly embedded markup which is often depreciated, provided that project the only efficient way of extracting these fragment with oll the contextual information that is necessary for a reader to make sense of them, in an dynamic edition that tries to avoid the presentational bias of the printed edition.

Bibliography:

- Robinson, Peter, "What text really is not and why editors have to learn to swim", Lit Linguist Computing (2009) 24(1): 41-52
- Schmidt, Desmond, "The inadequacy of embedded markup for cultural heritage texts", Lit Linguist Computing (2010) 25(4): 381-391
- Marilyn Deegan, Kathryn Sutherland, Transferred illusions: digital technology and the forms of print, Marilyn Deegan and Kathryn Sutherland (ed.), Ashgate publishing limited, England, 2009
- Buzzeti, Dino, "Digital Editions and Text Processing", in Text editing, print and the digital world, Marilyn Deegan and Kathryn Sutherland (ed.), Ashgate publishing limited, England, 2009.

A Register of Baroque and Enlightenment Slovenian Manuscripts: TEI encoded Analyses and Editions

Ogrin, Matija; Erjavec, Tomaž; Javoršek, Jan Jona

The Register of 17th and 18th century Slovenian manuscripts is a TEI (5) encoded archive of ms. descriptions and related digital facsimiles and is the first specialised digital collection of manuscript material in Slovenia, <http://nl.ijs.si:8080/fedora/get/nrss:nrss/VIEW/>.

In the paper, we give an outline of the content and structure of the register, and comment on some specific TEI encoding practices, used in the construction. The archive is focused on early-modern mss., esp. from the periods of baroque and enlightenment. In these periods, a great variety of textual genres and literary forms developed, where texts emerged in several, most distinct socio-cultural contexts, reaching from writings of civil and ecclesiastic persons to texts of peasants and self-educated writers. In our analyses, expressed through precisely structured ms. descriptions, we wanted to capture as many of these aspects as possible: not only codicological and historical data, but also formalised, searchable expressions for literary genres and socio-cultural background of particular ms. To this end, we prepared a taxonomy for textual genres and one for socio-cultural contexts, and linked the ms. descriptions to them. In this way, the msDesc element is not only a container for structured codicological description, but also a carrier of textual and socio-cultural analytical information.

Another aspect of the archive is the gradual preparation of transcriptions of selected manuscripts. In the process of editing, two transcriptions are prepared, a diplomatic one, and a separate critical (edited) text. Besides regular text-critical features, a problem specific

for mss. appeared: some portions of text in the source are mixed up and subsequently marked by the scribe in due order. In passages of this kind, the order of diplomatic and critical text differ substantially – but the encoding still has to allow for a parallel presentation of the transcriptions.

The descriptive, analytical and editorial perspectives to the manuscripts shed new light to each other in the form of a hermeneutic circle. We tried to grasp this semantics in the TEI encoding to unfold the complexities of the baroque and enlightenment manuscripts: unpublished and nearly unknown, but rich with research potential and aesthetic value.

Bibliography:

- Matija Ogrin (ed. by): Škofja Loka Passion Play. A Digital Critical Edition. Research Centre of Slovenian Academy of Sciences, 2009. <http://nl.ijs.si/e-zrc/sp/index-en.html>.

Creating, enhancing and analyzing TEI files: The new, XML-based version of TUSTEP

Ott, Wilhelm; Ott, Tobias

TXSTEP offers an interactive XML-based interface to the proven and powerful routines of TUSTEP, the Tübingen System of Text Processing programs. For more than 35 years TUSTEP is being developed and maintained at Tübingen University's computing centre. TUSTEP is a scripting language as well as a publishing system for the humanities, up until today unmatched in its overall performance and flexibility. TUSTEP primarily addresses users in the fields of the textprocessing humanities, such as computerlinguists, -philologists and editors. For more information, see www.tustep.org.

But, since its genuine syntax is proprietary, not intuitive and supposed to be difficult to learn, users tend to help themselves with other - often less effective - tools or less specific programming languages.

TXSTEP now gives a good answer to this situation by providing a user-friendly XML-syntax, allowing beginners and advanced programmers to use the whole scope of TUSTEP services in a modern, established programmers environment. The benefits are obvious: support of an open standard, widespread dissemination, programming in every other XML-editor, syntax highlighting, code completion and intelligible APIs. Moreover, TXSTEP is aided by the fact that there is no need to change the program's actual core. TUSTEP itself is open source as TXSTEP is going to be as well.

The purpose of TXSTEP, as well as of TUSTEP, is not to provide ready-made solutions for pre-defined problems. It "only" provides program modules for the basic functions of text analysis and processing.

It is the user who has to combine them in order to obtain the solution to a problem at hand. This is the prerequisite that he can take over the responsibility for every detail of the results obtained by computer application.

One of the features of TXSTEP is its capability to process almost all forms of textual data, whether this being XML-data or plain text files. Wherever there is textual data that has to be processed in the first place in order to gain TEI-data or to enhance the markup of insufficiently tagged XML data, TXSTEP is at its place.

The proposed demo is based on a prototype and shows the achieved state of our work in progress. It will demonstrate TXSTEPs functionality on the basis of tasks which cannot easily be performed by existing XML tools, including problems presented recently on the TEI list.

Bibliography:

- Digital publishing: tools and products In: Poiesis & Praxis: International Journal of Technology Assessment and Ethics of Science Vol. 5 Nr. 2 (2008) S. 81 – 112

The Role of Technology in Scholarly Editing

Pierazzo, Elena

In the past years two complementary but somewhat diverging tendencies have dominated the field of digital philology: the creation of models for analysis and encoding, such as the TEI, and the creation of tools or software to support the creation of digital editions for editing, publishing or both (Robinson 2005, Bozzi 2006).

These two tendencies are not necessarily mutually exclusive, as the creation of models can represent either the underlying structure or an exporting format for the development of tools. However, these two approaches have often been perceived in opposition, as a dichotomy. On the one hand we have the XML enthusiasts, the editors-as-encoders who apply XML markup to their texts and perhaps also develop publication strategies; on the other hand we have those who support out-of-the-box tools (the 'magic' or 'black' boxes), who proactively seek the development of fully comprehensive tools that present user-friendly interfaces with the explicit purpose of 'covering the wires', in particular hiding the much-abhorred angled brackets. But what are the implications of these positions with respect to the future development of digital (or computational) philology? How realistic is it to ask 'traditional' textual editors to turn into encoders? Conversely, how realistic and sustainable is the creation of 'magic boxes'?

In the past I have studied the difficulties and theoretical implications of using a TEI-based editorial model for an editorial team that was highly geographically dispersed (Pierazzo 2010, but presented as a paper in 2008). On that occasion I argued that the development of 'magic boxes' is a very ambitious item to have on the digital philology agenda because every edition, every scholar needs a very specialized, tailored set of tools. In the same article I expressed the opinion that, even if the scholars do not feel comfortable in using tags-on-view XML and the TEI, this was the only reasonable approach for digital scholarly editions. A couple of year later, my judgment has been mitigated somewhat. This was brought about largely by the interesting article by Tim McLoughlin (2010, to be read in combination with Rehbein 2010) which presents in an insightful way the difficulties and resistances in turning a consolidated editorial model into a digital TEI-based one, combined with the experience I

gained on some collaborative research projects at King's College London's Department of Digital Humanities: these together have triggered questions about the role of technology when it comes to digital scholarly editing. As a matter of fact, the evolution of the editor into an editor-encoder has yet to be investigated in full; at the moment it seems that the attention has been mostly devoted to the steep learning curve necessary to master the techniques of encoding in XML but without reflecting on the deep and sometimes unwelcome changes in the editorial work and workload once a new editorial model is undertaken, particularly when that model is based on TEI. This model sometimes sees the editor-as-encoder evolving also in the editor-as-programmer, the editor-as-web-designer and editor-as-(self-)publisher (Sutherland and Pierazzo 2011). These changes in the editorial work and role of the editors necessarily result in somewhat parallel changes in the final editorial products.

On the other hand the claim for the magic box seem to have receded somewhat, and we have witnessed the appearance of the interesting experience of creating configurable and standard-based tools that have the less ambitious goal of trying to help particular stages of the editorial work (collation, creation of stemmas and critical apparatus, transcription, annotation); this evolution is represented at best, in my opinion, by the tools developed within the Interedition (in particular with CollateX) and TextGrid projects.

This paper will briefly present the background outlined above, and then turn to fundamental issues that arise from it about the nature of editors and editing for digital editions. In particular, it will address the following questions:

1. Which are the competencies necessary for digital editors?
2. Which are the roles that digital editors are expected to cover?
3. What do editors expect the technology to do for them?
4. Which parts of the editors' work should be assisted by the computer and which must still be performed in the traditional way?
5. In which ways is digital editing different from traditional editing, if any?

Failing to understand how technology can really contribute to the editorial work will have serious consequences in the development and ultimately existence of digital editions.

The paper will address these theoretical and methodological questions making use of concrete examples, particularly from the Jane Austen Digital Edition and from the ongoing editorial experience of the Early English Laws project.

Bibliography:

- Bozzi, A. (2006). 'Electronic Publishing and Computational Philology'. In The Evolution of Texts: Confronting Stemmatological and Genetical Methods, C. Macé, P. Baret, A. Bozzi and L. Cignoni (eds.). Pisa-Roma Istituti Editorali e Poligrafici Internazionali.
- Pierazzo, E. (2010). 'Editorial Teamwork in a Digital Environment: The Edition of the Correspondence of Giacomo Puccini'. In Rehbein, M. and Ryder, S. (eds.). Jahrbuch für Computerphilologie, vol. 10, pp. 91-110. Also available at: computerphilologie.tu-darmstadt.de/jg08/pierazzo.html

- McLouglin, T. (2010). Bridging the Gap. In Rehbein, M. and Ryder, S. (eds.). Jahrbuch für Computerphilologie, vol. 10, pp. 37–54. Also available at: computerphilologie.tu-darmstadt.de/jg08/mclough.pdf
- Rehbein, M. (2010). 'The Transition from Classical to Digital Thinking. Reflections on Tim McLoughlin, James Barry and Collaborative Work'. In Rehbein, M. and Ryder, S., (eds). Jahrbuch für Computerphilologie, vol. 10, pp. 55–67. Also available at: computerphilologie.tu-darmstadt.de/jg08/rehbein.pdf
- Robinson, P. M. W. (2005). 'Current Issues in Making Digital Editions of Medieval exts ¬– or, Do Electronic Scholarly Editions Have a Future?'. Digital Medievalist, 1(1). Available at: www.digitalmedievalist.org/journal/1.1/robinson/
- Sutherland, K., and Pierazzo, E. (2011). The Author's Hand: from Page to Screen. In Deegan M., and McCarty W. (eds.), Collaborative Research in the Digital Humanities. Aldershot: Ashgate (forthcoming).
- CollateX: https://launchpad.net/collatex
- Early English Laws: www.earlyenglishlaws.ac.uk
- Interedition: www.interedition.eu
- Jane Austen Digital Edition: www.janeausten.ac.uk/index.html
- TextGrid: www.textgrid.de/

Reference and Annotation. From Citation to "Watson"

Prätor, Klaus

Reference, of course, is a crucial element of scholarly editions. It is the basis not only for classical external citation but also for editing itself: identifying, comparing, manipulating and annotating chunks of text – and also for programs that support or automate such tasks.

Markup has gone a long way from its beginnings to its nowaday use. Initially it was intended to encode THE logical structure of a document regardless of its later graphical form. Today, especially in editions, the markup has the task of preserving a multitude of interests for annotation, e.g. philological, linguistic, historical ones. Fundamental is the fact that with different and maybe evolving or changing interests for annotation the idea of a static, unchanging document, which was the concept of generalized markup at first, is vanishing. It is even questionable if it makes sense to conceive it as ONE document.

In this context an XPath-expression is no longer a sustainable reference. It may be changed by each later inserted tag. Especially programs for the retrieval or transformation make heavy use of the XPath. Realistically the text nowadays has to be seen as a work in progress. And in the present conception this work has the tendency to spoil the fundaments of its own reference system.

In my opinion a solution consists of two parts:

One is to refer only to relatively stable parts of the structure of a document, essential meaningful elements, leaving aside the tagging of annotations etc. In many texts this elements could be paragraphs, sentences and words. These, or appropriate similar ones for special text species, should be the base as well for external citation as for the internal, manual or programmed, editing of the text.

The other thing is considering a radically new organisation of annotation. A fundamental different approach to representing annotations as inline markup is referred to as the “stand-off” annotation model. In a “stand-off” annotation model, annotations are represented as objects of a domain model that “point into” elements of the unstructured content) rather than as inserted tags that affect and/or are constrained by the original form of the content.[UIMA]

There are undeniable advantages, especially in the combination with the suggested basic markup of meaningful elements. Firstly these may serve as elements the stand-off markup can refer to. The next strong point is that annotation can be handled and searched independently of the central document. Furthermore, within the stand-off markup there is no need for overlapping and finally the combination of a linguistically organized basic structure and separate supplementary metadata is a perfect basis for a semantic approach to texts.

This has not gone unnoticed. TEI acknowledges the role of stand-off markup and also other authors, including Robinson, Sahle and myself, are mentioning a potential need for stand-off markup in recent papers. Most remarkably there exists already a standard for this sort of annotation, the Unstructured Information Management Architecture (UIMA). Its Common Analysis Structure (CAS) consists of two fundamental types of objects

- Sofa, or subject of analysis, which holds the artifact (the original document)
- Annotation, a type of artifact metadata that points to a region within a Sofa an “annotates” (labels) the designated region in the artifact.

Wherever possible UIMA is following established standards. It formulates no domain specific models and so could be complemented by conventions e.g. of TEI.

The potential of this approach has been shown in a popular, nonetheless impressive example by IBM. Its Deep Question Answering System (called “Watson”) is based on UIMA and was able to beat in a prominent US Quiz-Show (Jeopardy!) two human champions. Aside of UIMA DeepQA is using a system of shallow semantic parsing of natural language documents. While UIMA is implemented by Apache mainly in Java (but open for other programming languages), the parsing and also the handling of the database in Watson is done in Prolog. UIMA documents can be transformed in inline markup and/or in RDF. This can also be done selectively, producing individual representations or views of the original document together with its metadata.

In one respect the implementation of UIMA differs from the idea suggested in this paper. In UIMA an annotation points simply to some character string as region of reference in the original document, while in the approach of this paper the metadata point to a list of symbolic elements (words or even sentences). From a theoretical point of view this seems preferable. If this is true also in practice, only practice can show. Therefore it is a pleasure that we could start an implementation of these ideas within the digital part of the edition of the work of Jean Paul in Würzburg.

Bibliography:

- Zur Zukunft des Zitierens. Identität, Referenz und Granularität digitaler Dokumente (erscheint in Editio)
- Ceci n'est pas un texte? Zur Rede über die Materialität von Texten - insbesondere in den Zeiten ihrer Digitalisierung, in: Martin Schubert, Materialität in der Editionswissenschaft, Berlin 2010 (Beihefte zu editio)
- A Model for Memory. Synergies in Sparse Matrices, in: Klaus Mainzer ECAP10. VIII European Conference on Computing and Philosophy, München 2010
- Individuen und Referenzobjekte. In: Methodisches Denken im Kontext. Hrsg. von Peter Bernhard und Volker Peckhaus. Paderborn (mentis) 2008
- Topologie und Navigation. Zur Bewegung in elektronischen Editionen, in: Editonen - Wandel und Wirkung, hrsg. v. Annette Sell, Berlin 2007
- Kollationen und Transformationen für XML-Dokumente (mit Dietmar Seipel), Berlin 200
- XML Transformations Based on Logic Programming (with Dietmar Seipel), in: 19th Workshop on (Constraint) Logic Programming, Ulm 2005
- Logic for Critical Editions, in: Proceedings of the 15th Int. Conference on Applications of Declarative Programming, München 2004

Logging the Abbot: Reflection-Oriented XSLT Programming for Corpora Conversion and Verification

Pytlik Zillig, Brian L.

It is a substantial challenge of digital text curation that similar but distinct collections sometimes must be made to interoperate by a lossless conversion into a common format such as TEI. While it can be relatively easy to verify losslessness for a small text collection, it becomes more difficult with more texts. Curation and verification routines that rely on individual human scrutiny will not operate at a large scale or in a reasonable amount of time.

In early 2007, the MONK Project began to develop a procedure for batch-converting varying collections of XML-encoded texts into a specialized application of TEI P5 that we called TEI-Analytics (TEI-A). The effort to develop a conversion procedure yielded a command-line application, which was called Abbot. Abbot works by analyzing the XML schema that describes the document structure to which the target collection should be converted. Abbot then uses that analysis--an enumeration of allowable elements and their associated attributes--to programmatically generate an XSLT stylesheet that is used for the conversion.

By mid-2009, Abbot had successfully converted 2,585 texts, all of which were valid according to the TEI-A schema. It is a well-known fact, however, that a text collection in possession of document validity may still be in want of some additional form of scrutiny, if only to verify that no words were inadvertently lost or rendered out of sequence. For MONK and its texts, this scrutiny was undertaken on a selective basis by members of the project team. This approach worked for the MONK texts. In the case of the roughly 30,000 texts produced by the Text Creation Partnership--a collection more than ten times larger than the MONK corpus--the problem of validating markup and verifying textual fidelity becomes clear and new procedures are needed. Distant reading, according to Moretti, is the sort of reading that one does when there are too many texts to read closely. Similarly, "distant verification"

becomes necessary when there are too many text alterations, or too many texts, to verify closely and individually.

As a developer (with Stephen Ramsay and Martin Mueller) of the original Abbot software, I have extended Abbot to be able to verify the fidelity of all transformations by measuring the inputs and outputs, and calculating and logging every difference. For each XML node, a log entry is made that records any changes to the node, including: (1) the node name, (2) the names of child nodes, (3) the attribute names and values, (4) the text nodes that are children of the current node, and (5) counts of each of the above. All changes are, by default, made as part of the Abbot transformation pipeline and logged in a file that is produced in comma-separated-values (CSV) format. While a command-line diff operation could potentially be used to perform the task of comparing XML files to their source texts, Abbot adds this functionality as a first-class operation to the processing pipeline. The CSV format makes it a trivial task for a spreadsheet program to calculate the consequence of a given conversion, or all conversions.

Every substantive change to the XML structure or to the text content is recorded. Abbot's measurement of nodal difference is not based on simple string comparison, which would report trivial differences such as <foo n="1" id="a"/> and <foo id="a" n="1" />. In this example, the order of the attributes is reversed, but the two nodes are otherwise the same. XML differencing applications do exist, but are not sufficient, because they are not able to refer to the functions or templates responsible for a given change. The same pipeline that alters the XML input nodes and writes the output nodes should be able--as Abbot now is--to log all differences.

For Abbot, logging is enabled in each XSLT template and each template is reflective. Templates are created at runtime based on input that is compiled at runtime. They vary depending on the source texts and on the desired output schema (TEI-A, TEI-All, or something else). Reflection in this context produces templates that are advantageous in several ways: they are self-identifying, self-describing, and self-differencing. The log file records the location of every change, a description of the alteration, a quantitative measurement, and the unique ID of each template responsible for the adjustment(s).

The change-logging extension of Abbot, by making the integrity of texts verifiable across transformations, solves an important obstacle to keeping curated data meaningful. Gold asserts that a "great challenge of data curation is ensuring that data, once preserved, remains meaningful either within the same research area or ideally across areas or even across domains." When the happy day arrives, perhaps soon, that we have at our disposal the "million[s of] books" that Crane writes about, we will curate them with precision and care and caution and a complete accounting of alterations.

Bibliography:

- Crane, G. "What Do You Do with a Million Books?" D-Lib Magazine, Vol. 12, No. 3, March 2006.
- Gold, A. "Data Curation and Libraries: Short-Term Developments, Long-Term Prospects." Library, California Polytechnic State University, San Luis Obispo. April 4, 2010. Retrieved May

13, 2011, from digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1027&context=lib_dean
- Moretti, F. Graphs, Maps, Trees: Abstract Models for a Literary History. New York: Verso, 2005.
- Pytlik Zillig, B.L. "TEI Analytics: converting documents into a TEI format for cross-collection text analysis." Literary and Linguistic Computing, Vol. 24, No. 2, 2009, pp. 187-192.

Realistic targets in TEI to RDF

Rahtz, Sebastian

It has been a target of the TEI Ontology SIG for some time to work out mappings between TEI elements and RDF vocabularies. Most notably, there has been a concerted effort to align the TEI with ISO 21127:2006, the CIDOC conceptual reference model. The aim of this paper is to review how far this can be implemented in practice, with the aim of taking an arbitrary TEI document and extracting useful RDF assertions in CIDOC CRM, and thus enabling TEI digital texts to participate more fully in the world of open linked data. There are three aspects to this work:

- how to record the relationship of TEI elements to known CIDOC CRM concepts in a formal way, maintained in a single document with the mapping guidelines
- how to write a application to take the mapping specification and get from TEI XML to RDF XML
- how to embed the work in a system such as OxGarage to allow users to submit a TEI file to a web service and get back an RDF file

We may note that this work is in contrast to linguistic analysis of text, and extraction of assertions from published literature using NLP.

The background to the current effort is the CLAROS project at Oxford which aims to combine discrete databases of information about the ancient world using an RDF triplestore of assertions using CIDOC CRM. It includes art objects, archaeological sites, antiquarian photographs, and onomastics, and the latter data comes from the "Lexicon of Greek Personal Names" which is available as TEI XML. Getting LGPN include CLAROS involved setting up a TEI to CRM workflow.

A basic tool for recording mapping is provided by the TEI in the form of the <equiv> element, which allows us to provide a specification in ODD which points from a TEI element to an external identifier, and says how to get there. This allows us to build a sensible extraction tool. There are many ambiguities and details to resolve along the way, especially when we work with less structured, but richly marked-up, text. Among the problems one encounters are

- how to record the location in the TEI text of an RDF assertion
- how to provide metadata (eg date and author of the assertion)
- the representation of uncertainty and precision
- what conventions to adopt for chronological periods,

spatial coordinates and dates, the precise expression of which are left vague in CIDOC CRM. The paper gives examples of the results obtained from a variety of TEI texts, demonstrates the implementation in OxGarage of a useable converter, and shows how the resulting RDF can be queried.

A web-based application for rapid annotation of TEI documents

Ritter, Jörg; Andert, Martin; Molitor, Paul

Annotating literary texts based on the recommendations of the TEI [1] can become very cumbersome when editing XML files directly. The double-end-point attachment method, e.g., requires in-depth knowledge of XML and introduces a lot of markup. The basic task of creating an annotation however is very similar to formatting text in word processing programs, where it is convenient to just select some text with the mouse in order to change its appearance or take further action on it. In this paper we present a web-based tool for rapid annotation of TEI documents by just selecting passages of text in a purpose-made preview.

Crafting the markup of passages of text together with links to comments and annotations is a tedious and error-prone process when done directly in XML. Even sophisticated XML editors like oXygen [2] with support for unique XML identifiers, tag and attribute auto-completion fail in making this procedure a more pleasant experience. One of the proposed methods of the TEI guidelines – the double-end-point attachment method – requires two XML tags (attributed with unique identifiers) to indicate the start and end of a lemma. The annotation itself – contained in another tag – is then linked to the lemma using these identifiers. The built-in annotation method of our system completely hides this rather complex workflow from the user by providing several tools that mimic the behavior of annotating a printed text in reality using pens and labels in different colors and sizes. In doing so it is very similar to the above mentioned, comfortable, and widely-used method of formatting text in word processing programs: the user selects a phrase with the mouse pointer and then takes further action on it like boldening or coloring it differently. Provided that we have an adequate preview of the XML, the user just selects a text snippet in the preview and the annotation is created automatically in the background using the sophisticated methods proposed by the TEI guidelines. Similar approaches are well-known with respect to PDF documents and web sites [3,4].

Following this idea we built a proof-of-concept application for rapid annotation of TEI documents. This web-based application provides easy-to-use tools for searching, highlighting and commenting of passages of TEI encoded text. Thus it is applicable to researchers who might have little or no knowledge of XML and the TEI guidelines. Given a TEI document to be annotated we provide a suitable preview. While exploring the preview or using our infix search we offer virtual felt-tip pens and adhesive labels for annotation. Beside the standard highlighter the user can create additional ones with customized names and colors. All annotations are listed in tabular form next to the preview and are connected with the corresponding positions in the preview through linking. Behind the scenes each modification by the user is sent to the server where the XML processing occurs. At any time the user can export the annotations themselves, a preview of the underlying document with or without

the annotations, or the XML file enriched with TEI conformant annotations. At no time the user has to manually insert the required TEI tags and their attributes nor edit raw XML code by hand.

The presented application is basically a JavaScript enhanced and database driven web site, so there is no need for the end-user to manually install desktop software on his local computer; a working web browser that is connected to the internet is the only requirement. The application has been tested to work with the Firefox browser in version 3.5 or higher. Because Firefox is available for all major operating systems, our application is platform independent by nature. The application itself exhibits a number of special features that aid the user in annotating a text. Besides having a facility to upload and save a file containing a TEI conformant XML document, our application generates a preview of the uploaded TEI document utilizing XPATH and XSLT. Of course, in order to obtain an annotatable preview both the XPATH expressions and the XSL transformations used have to be fine-tuned for the specific kind of text at hand. The implementation of such expressions/transformations is beyond the scope of this article and will be presented in another, more technical paper.. As a proof of concept our system supports TEI encoded performance texts (according to P5 Guidelines, Chapter 7). Upon highlighting and commenting a phrase in the text, the web browser makes an AJAX-based round trip to the server taking only the modified fragment of the HTML preview with it. Using a reverse XSL transformation, the server integrates the user's modification into the source TEI file and saves it if necessary. The client then updates the preview display and the corresponding annotations list.

We have evaluated this approach on prose and performance texts. Enclosed please find a screenshot of this proof of concept.

Bibliography:

[1] tei-c.org
[2] oxygenxml.com
[3] www.adobe.com/products/acrobat.html
[4] www.awesomehighlighter.com/

The Descartes Corpus (ProDescartes, ANR 2009-2013) Presentation

Roger, Julia

The communication will consist in the presentation of The Descartes Corpus Project which aims at an online edition of all the works and correspondence of Descartes. It is led by the team "Identité et Subjectivité" from the University of Caen (Basse-Normandie), in scientific collaboration with the "Centro interdipartimentale di studi su Descartes e il Seicento" from the University del Salento (Lecce), the Centre d'études cartésiennes" (from Paris IV) and the GREYC (Caen) ; and in editorial collaboration with the Presses universitaires de Caen and the Bibliothèque nationale de France (BNF) .

This project, under development, has these main objectives :

1. to publish online and in text mode (XML-TEI) the original editions of the works and letters (the Clerselier edition according to the copy of the Institute of Correspondence, with the transcription of the original footnotes and bookmarks) as well as the scanned pictures of the pages from the original edition and the Adam-Tannery edition ;
2. to develop and integrate the corpus a scientific annotation tool that would take advantage of the digital edition : it would make it possible for the scientific editors to create and especially to modify their own "footnotes" online, instantly via the Web interface, after the scientific editor's approval ;
3. to develop a full-text and trilingue search engine browsing TEI-XML files (to give occurrences of the searched word both in contemporary and classical French language and in Latin) ;
4. to offer the readers a reservoir of studies or technical notes about Descartes's publications in the form of articles, reports or more important works. These studies would be on the sidelines of the cartesian corpus and available from the website home page, after the scientific editors' approval.

The communication will take place in these topics of reflexion :

- The relation between representation (encoded text) and presentation (visualisation, user-interface, points 2 and 3) ;
- TEI encoded data in the context of quantitative text analysis (point 3) ;
- Integrating the TEI with other technologies and standards (points 1, 2, 3) ;
- TEI as interchange format: sharing, mapping, and migrating data (in particular in relation to other formats or software environments) (point 3).

Bibliography:

- J. Roger, « Présentation du projet Corpus Descartes (ProDescartes, ANR 2009-2013) », Le Patrimoine à l'ère du numérique (Actes du colloque international « Le patrimoine à l'ère du numérique : structuration et balisage », Université de Caen, MRSH, à paraître ;
- J. Roger, « "Corrompere il senso" ». Gli a capo nella Seconda Meditazione, Alvearium, 2011, IV, sous presse.

The Critical Step in Open Content Greek: Towards a Digital Edition of Athenaeus

Romanello, Matteo; Berra, Aurélien

Collaboratively building a comprehensive library of linked-up Greek classical texts has now become part of our digital horizon. Any such prospect depends on versatile and well-accepted standards. Thus, *The First Thousand Years of Greek* is a project which aims to provide TEI-compliant, morphologically-tagged versions of most Greek texts from Homer to the imperial age [1]. To be truly naturalised in scholarly practices, such libraries will have to integrate the critical dimension which major digital collections like the *Thesaurus Linguae Graecae* have set aside so far. Interestingly, the latest release of XML texts in the Perseus Project was presented as a kind of potlatch: classicists should now take up the challenge of using and improving this offering of "Open Content Greek" [2]. Athenaeus' compilation

being one of these texts, our reflections may be seen as a response to this important initiative.

What should be a new edition designed with an awareness of this transforming landscape? Although *Digital Athenaeus*, our nascent project, is connected to a traditional philological undertaking, we engage in a natively digital editing process. In our view, this material both demands and rewards new approaches. Indeed, we think of this *editio princeps electronica* as an experiment at the interface between breadth (large-scale, extensive work) and depth (deep encoding, intensive philological work) [cf. 3,4].

In this paper we do not intend to present a complete rationale. Instead, we want to discuss the suitability of TEI encoding in this classical case. In the first part of the paper we deal with the implications of using text mark-up to represent its structural features and to devise a critical apparatus. In the second part we consider those aspects of an edition for which the combination of TEI with conceptual models can be a viable solution, in particular the alignment of multiple text versions and the formalisation of their relationships.

The Deipnosophists, or *Learned Banqueters*, is first and foremost an erudite digest: Athenaeus, a Greek author from Egypt who was active in Rome around 200 CE, wrote a gigantic miscellany of texts pertaining to the alimentary and cultural components of the symposium. In the course of some 1,500 pages and 300,000 words in the reference Teubner edition, he adduces thousands of quotations, sometimes famous and preserved in fuller versions, often utterly obscure and otherwise lost. The opacity created by this superabundant matter contributed to Athenaeus' status, that of an ancillary witness which one frequents somewhat resentfully.

In their loosely thematic structure, the fifteen books follow the parts and practices of a banquet. They combine several layers of dialogue and narration, since Athenaeus reports and comments on the meetings of more than twenty characters. At every level, they extend a nexus of quotations from over 1,200 authors. The establishment of the text would be rather simple — with one 9th-10th c., almost complete manuscript of the unabridged version and apographs —, were it not for the existence of a 12th c. epitome whose textual origin remains unresolved. Given its importance as a source for classicists and historians, the work has had several editions, translations and commentaries; due to its daunting bulk and unusual subjects, this amount of scholarly material remains manageable.

In the long run, the aim of our digital project will be to offer the whole Athenaeus dossier in the form of an evolutive virtual research environment. As the editors will be among its first users, it might also be termed a virtual editing environment.

It is essential to make visible the structure by carefully encoding three overlapping structures whose distribution has not been adequately studied: the characters responsible for the speeches and quotations, the topics treated, the authors and works quoted (quotation, paraphrase and allusion should be distinguished, as well as references to extant and lost works). The bearing of context on the interpretation of fragments makes it a requisite. Furthermore, Athenaeus' reflexivity is largely embedded in the organisation of internal

references, comments and sequences of quotations. To enable a valid perspective, we need accurate reading tools.

The elaboration of a critical apparatus raises interesting issues. How to record differences in the frame of a full-blown environment proposing several editions? This is not the same problem as constructing a digital surrogate for printed apparatuses (on which process [5] can be consulted). Even if we only contemplate the preparation of a new eclectic text, the full collation of the witnesses shall give a clearer view of this tradition: automatic comparisons and statistical data should contribute to the act of editing. They may take us beyond the prejudice against or in favour of Byzantine capacities for conjectures, which is still predominant in the debates on the relationships between the unabridged and the epitomised version.

Thus, we are aware of the potential of handling the “many texts” (incidentally, unrestricted databases of variants are also databases of revealing manuscript mistakes and meaningful erroneous conjectures). Nevertheless, we think that it is crucial to provide the “one text” which many readers want to use as a reference. And its trustworthiness is increased by the adoption of the “Open Source Critical Editions” model, which implies a reassessment of what a critical edition is, in relation to our current capabilities in terms of accessibility, transparency and explicitness. (On the “one text” and “many texts”, see [6]; on the OSCE model, see [7]).

Hence, the understanding of Athenaeus’ work vitally depends on contextualisation and is inherently a matter of structure, while the issue of transcription challenges more generally the received notion of critical method. This leads us to read differently a statement by John Lavagnino [8]: “Two of [the requirements of a TEI approach] can be especially problematic: first, you need to understand your texts; second, you need to believe in the integrity and utility of selective transcription.” While Lavagnino referred to stages in a project or cases when TEI should not be used (e.g. when Encoded Archival Description or DocBook formats are more adequate), for other aspects of the project we might want to combine a TEI schema with other frameworks.

What are the other technical solutions from which a TEI-based “virtual editing environment” could benefit? There have been some recent and convincing attempts both to mix a TEI layer with an ontological approach [9] and to align TEI elements with classes and properties defined in ontologies such as the Conceptual Reference Model of the International Committee for Documentation (CIDOC-CRM) [10].

The main point we want to make in support of this method is that building a collection of digital texts upon a well-defined conceptual model can prevent us from creating misleading representations, as it was argued by [11,12]. A digital collection which contains both texts entirely preserved in manuscripts and literary fragments preserved only as embedded quotations, for example, will inevitably contain duplicates unless the underlying model defines and implements the concept of citation. Duplicate records will otherwise alter the results of any quantitative analysis, such as word frequency. Given the number of quotations contained in *The Deipnosophists*, this requirement becomes pivotal for us: we want to be

able to isolate the quotations from the rest of the text, whether for quantitative or qualitative analyses.

Within the architecture of our project TEI encoding is to be combined with the use of the Canonical Text Services (CTS) protocol, on the one hand, and with CIDOC-CRM and Functional Requirements for Bibliographic Record Object Oriented (FRBRoo), on the other. The reason for this choice is also to separate the mark-up of the text, and of each of its "versions", from the definition of the relationships that exist between those different versions. By versions of the text we mean, for example, the diplomatic transcriptions of the extant manuscripts, or the available critical editions. The CTS protocol is specifically used to align with each other these versions, whereas CIDOC-CRM and FRBRoo allow us to define their relationships (X is an edition of Y, Z is a translation of X, K summarises Y, and so forth).

Finally, the use of an ontology-based layer aims to make more explicit and machine-understandable the statements usually implicit in critical editions. For instance, through a diagram like the *stemma codicum*, the editor intends to formulate a hypothesis about the history of the text considered. It should be noted that there have been interesting attempts at expressing and encoding such statements by using an event-based ontology like CIDOC-CRM. Such ways of augmenting the TEI might be instrumental in bringing digital editions to their full potential. Really digital and fully critical editions remain a desideratum in the classical field.

Bibliography:

[1] Center for Hellenic Studies. The Free First Thousand Years of Greek, a project directed by Neel Smith. Harvard University, 2008-. chs75.chs.harvard.edu/projects/diginc/first1kyears.

[2] Crane, Gregory. "Plutarch, Athenaeus, Elegy and Iambus, the Greek Anthology, Lucian and the Scaife Digital Library – 1.6 million words of Open Content Greek." The Stoa Consortium, December 13, 2010. www.stoa.org/archives/1332.

[3] Crane, Gregory. "Give Us Editors! Re-inventing the Edition and Re-thinking the Humanities." Paper presented at the conference The Shape of Things to Come, Charlottesville: 2010. shapeofthings.org/papers/ (draft).

[4] Boschetti, Federico. "Digital Aeschylus. Breadth and Depth Issues in Digital Libraries." In AT4DL 2009. Workshop on Advanced Technologies for Digital Libraries 2009, 5-8. Trento: 2009. www.unibz.it/en/public/universitypress/publications/all/Documents/9788860460301.pdf.

[5] Mastronarde, Donald J. "Towards a New Edition of the Scholia to Euripides." Paper presented at the American Philological Association Conference, 2008. Available in his Euripides Scholia Online Edition. euripidesscholia.org/EurSchHome.html.

[6] Robinson, Peter M. "The One Text and the Many Texts." Literary and Linguistic Computing 15.1 (Special Issue "Making Texts for the Next Century", 2000): 5-14. llc.oxfordjournals.org/content/15/1/5.abstract.

[7] Bodard, Gabriel and Juan Garcés. “Open Source Critical Editions: a Rationale.” In Text Editing, Print and the Digital World, edited by Marilyn Deegan and Kathryn Sutherland, 83-98. Farnham: Ashgate, 2009.

[8] Lavagnino, John. “When Not to Use TEI.” In Electronic Textual Editing, edited by Lou Burnard, Katherine O’Brien O’Keeffe and John Unsworth, 334-338. New York: Modern Language Association of America, 2006. www.tei-c.org/About/Archive_new/ETE/Preview/lavagnino.xml (preview version).

[9] Ciula, Arianna, Paul Spence and José Miguel Vieira. “Expressing Complex Associations in Medieval Historical Documents: the Henry III Fine Rolls Project.” Literary and Linguistic Computing 23.3 (2008): 311-325. llc.oxfordjournals.org/content/23/3/311.abstract.

[10] Ore, Christian-Emil and Øyvind Eide. “TEI and Cultural Heritage Ontologies: Exchange of Information?” Literary and Linguistic Computing 24.2 (2009): 161-172. llc.oxfordjournals.org/content/24/2/161.abstract.

[11] Romanello, Matteo, Monica Berti, Federico Boschetti, Alison Babeu and Gregory Crane. “Rethinking Critical Editions of Fragmentary Texts by Ontologies.” In Proceedings of the 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies, edited by Susanna Mornati and Turid Hedlund, 155-174. Milano: 2009. conferences.elpub.net/index.php/elpub/elpub2009/paper/view/158/66.

[12] Romanello, Matteo. The Digital Critical Edition of Fragments: Theoretical Problems and Technical Solution. 2011. eprints.rclis.org/handle/10760/15592 (pre-print).

TEI and DARIAH: Current Activities and Future Work

Schöch, Christof; Volkmann, Armin

This paper is concerned with the relation between the TEI and DARIAH (Digital Research Infrastructure for the Arts and Humanities). These two endeavors are not only based on partially overlapping research communities, they are also acting in a shared context of the digital humanities, where they are participating in some of the major trends and issues. Therefore, analyzing the current areas of overlap and the potential for future interaction may prove to be of interest for the development of both DARIAH and the TEI.

In order to provide some context, we start by laying out the general aims and tasks as well as the organizational, disciplinary and thematic structure of DARIAH. This EU-funded, large-scale project is an international consortium “aiming to enhance and support digitally-enabled research across the humanities and arts” (DARIAH Mission Statement). This aim is pursued in four virtual competency centers which are concerned, respectively, with building domain-specific research infrastructures, fostering digital humanities research and education, developing standards and recommendations for research data, and developing strategies in the area of advocacy and outreach for digital humanities.

It is against this background that the relationship between the TEI and DARIAH will be considered, the TEI being viewed not only as a (de facto) standard for textual data, but also as an institution and a community. When looking at TEI and DARIAH in this way, the number of issues at hand is obviously vast. In the main part of the paper, we would like to focus on three issues that seem particularly relevant: standards and metadata, tool development, and community building. For practical reasons, we will focus on examples from the German contribution to DARIAH, but many of the issues concern DARIAH more generally.

Our analysis is careful to consider, at this still relatively early stage of the DARIAH project, the role the TEI is already playing in DARIAH as well as the future perspectives for the TEI in this project. At the same time, we look not only at areas in which the TEI can serve as a model to the DARIAH project, but also and quite specifically at ways in which DARIAH can contribute impulses to the further development and dynamics of the TEI. In this way, we hope to offer some insight into existing areas of common ground, but also to reflect on further areas of possible cooperation, the consideration of which may help foster the development of both TEI and DARIAH.

Micropapers

Digital Editions as the Myth of Sisyphus

Burghart, Marjorie

The Myth of Sisyphus is well known, even to schoolboys fascinated by Greek mythology: for having defied the gods and put Death in chains, Sisyphus was condemned to push a huge boulder up a mountain slope, and when he reached the top, the boulder would roll down the mountain, and Sysiphus would have to roll it up, again and again, for all eternity. In this paper, I will argue that this myth functions as an interesting allegory of the work of digital editors working with the TEI. When they decide to create a digital edition, they defy the limitations of traditional print editions, and give their work better chances for accessibility and perennity. But like Sisyphus, they have to pay a price for that: when a scholar publishes a print edition, he goes through the usual process of a critical edition, prepares his work for publication, then once it is published and printed, he does not need to care about his work: once it has been published, it will never be necessary to search for funds, unless a second edition is planned; libraries all over the world will take care of keeping copies safe in their collections and making them reasonably available to potential readers; the layout and typography of the work benefits from a secular tradition and is not likely to be questioned during the lifetime of the scholar.

In a word, the critical editor can move on to his next work, without giving much thought to his published edition. The scholar who engages in a digital edition, on the other hand, is a modern Sisyphus: publishing a digital edition is an exacting, never ending task! For instance:

- digital editions may often be at the mercy of the whim of IT services, willing or not to offer the necessary framework and evolutive development – this regards the encoding of the edition (with evolutions of the Text Encoding Initiative) as well as the applications needed to make an edition available to its readers
- the standards and trends of web design evolve very quickly, more quickly even than the technologies behind, and will demand a graphic overhaul every 3 or 4 years.
- to a lesser extent, the very possibility of updating the edition is a form of pressure in itself.

I will review these issues, showing how scholars publishing digital editions find themselves in an absurd situation where nothing guarantees their work will always be available, unless, during their whole career, they take care of maintaining their work – not to mention the uncertain future of their work once they retire.

I will discuss potential ways to address these issues and to relieve the critical editor of his Sisyphean task, among which I will suggest a better defined status of published digital editions, and the creation of public institutions offering the equivalent of a legal deposit to digital editions.

Bibliography:

- Hans Walter Gabler, “Theorizing the Digital Scholarly Edition”. Literature Compass 7/2 (2010), 43-56

Converting legacy editions to TEI

Cramme, Stefan

In the course of the last years, the Library for Research on Educational History (Bibliothek für Bildungsgeschichtliche Forschung, BBF) in Berlin, Germany, has offered scholars in the history of education a service to put online unpublished editions of sources which have been transcribed and prepared but not published yet (in contrast to the more numerous editions born digital or the retrodigitization of older works). These editions, ranging from the 18th to the 20th century, mostly have been begun with a conventional printed publication in mind which has not proved feasible, though. The texts are converted from legacy formats (usually an older version of Microsoft Word) to XML, applying a limited set of semantic TEI markup. The online editions generated from the XML version are in some cases accompanied by a printed volume with selections from the full corpus. As a research library, the BBF regards close cooperation with the research community as one of its main tasks, providing TEI expertise while the scholars can concentrate on revising the text and preparing supplementary material like indexes and annotations.

The micropaper will show the specific problems and pitfalls encountered in such a conversion process, but also focus on how unpublished legacy transcriptions and editions of source material can be adapted to the TEI guidelines and put online with limited resources. Examples will be taken from the correspondences of the educator Friedrich Fröbel and the educational philosopher Eduard Spranger.

Bibliography:

- Editionen in einer bildungshistorischen Forschungsbibliothek. In: Bibliothek und Wissenschaft, forthcoming summer 2011

Application of TEI to a biographical dictionary (www.Deutsche-Biographie.de)

Reinert, Matthias

Starting point of our DFG-funded project (2007-2009) had been two series of digitized volumes. The „Allgemeine Deutsche Biographie“ in 55 volumes (ADB, publ. 1875-1912) and 23 of „Neue Deutsche Biographie“ (NDB, publ. since 1953) comprising some 47.000 articles and 88.000 persons mentioned in a separate non-XML database. The raw text had typographical encoded features we used to automatically restructure the text. Each article (in NDB) consists of functional parts like genealogy, life, works, etc. Most challenging part was the realignment of articles to persons. We had to identify persons with biographies and those mentioned in the text. We did heavily use <persName>, <birth> and <death>-tags to

identify strings as persons. The TEI-Lite standard was released in order to allow those tags within text. While proof-reading the automatically encoded articles these tags could be more easily be hold apart than <name>-tags with several types. Nonetheless the TEI-Lite scheme can be validated after a simple transformation. In addition abbreviations and short-titles had been identified. Almost all persons mentioned in both series are manually identified in bibliographic Authority Files (Personennamendatei, PND). Places of birth and death are in progress to be aligned with geodatabases (namely OpenStreetMap). Both identifications result in concordance files, partly to have them maintained separately, partly to reduce the code within the TEI-encoded texts to ease readability. Next steps consists in a) „parsing“ the genealogy, to make relations between persons mentioned more explicit, b) breaking up the series/volume-structure, to have articles ordered and editable by person.

Bibliography:

- "Biographisches Wissen auf einen Klick", in; Akademie Aktuell, Heft 04/2010, S. 44-46 <http://www.badw.de/aktuell/akademie_aktuell/2010/heft4/16_reinert.pdf>

Mapping metadata of TEI-encoded biographies to CIDOC-CRM

Reinert, Matthias; Riechert, Thomas

While publishing online two digitised biographical dictionaries containing biographies for about 40.000 historical persons in 47.000 articles a major challenge is to make the data available. Beside presenting the material freely available online and porting metadata into academic search engines and OA-registries we choosed to create Linked Open Data out of our biographical repository. Funded by PUBLINK (part of LOD2.eu) the AKSW helped us to provide biographical metadata in RDF. Thanks to having almost all persons aligned with the German Name Authority File (PND, already part of LOD), adding to a majority of places of birth and death identified in Geodatabases (OpenStreetMap) we created a first set of common ontologies (FOAF, DCMES) to express statements like „was born in“, „died in“, „knows“. In a second step we defined a set of mapping rules to CIDOC-CRM (actually using the OWL-DL variant Erlangen CRM). Motivation has been

- to maintain easy interoperability with Europeana and the emerging Deutsche Digitale Bibliothek (German Digital Library)
- to be able to use semantic wikis (Wiss-KI.eu, Ontowiki.net, SMW+ recently in evaluation) assisting the redaction and correction of our content as well as „content-enrichment“.

Bibliography:

- Biographisches Wissen auf einen Klick, in: Akademie Aktuell, 4(2010), S. 44-46 <http://www.badw.de/aktuell/akademie_aktuell/2010/heft4/16_reinert.pdf>

Virtual Scriptorium St. Matthias

Vanscheidt, Philipp; Scholzen, Sabine

The project Virtual Scriptorium St. Matthias intends to reunite the worldwide scattered codices from the library of the Benedictine abbey St. Eucharius or St. Matthias in Trier electronically. The project is realized at Stadtbibliothek and Stadtarchiv Trier as well as at the Center for Digital Humanities at the University Trier since summer 2010. About 450 codices from the period between the eights and fifteenth century will be digitized in three years. These codices concern a wide range of topics from various traditions. Beyond theological and religious writings you find a large amount of latin classics like Cicero, Priscian, Sallust or Martianus Capella. A prestigious example for the inculturation of ancient and pagan spirit is an illustrated edition of Aesops fables. No other abbey possessed as many manuscripts of Hildegard of Bingen as St. Matthias. You find also three important specimina of Dectretum Gratiani. One of them includes 60% of all glosses ever written on this work. But the richly illustrated Trierer Apokalypse from carolingian times maybe the most famous of all these codices. The project Virtual Scriptorium St. Matthias will present an electronic catalogue that sums up the knowledge from older descriptions and combines them with a presentation of the digitized codices. In this context TEI is used as a standard of XML description of manuscripts. The amount of objects requires a synchronization of these descriptions with a dynamic database to correlate them with other digitized catalogues, editions and databases like the PND. The results will be integrated in Manuscripta Mediaevalia and TextGrid. In this way the project will not only provide images and metadata but will also be included into a virtual working space in where further research and exploring will be possible, e.g. with TEI concurrent transcriptions of selected works.The project homepage www.stmatthias.uni-trier.de will be released on the first of August 2011 on a trial base. The project should be presented in a short talk and a poster. The poster will cover the project thoroughly while the short talk is supposed to sketch the advantages and some practical limits of TEI in such an enterprise.

Bibliography:

- Geschichte in Metaphern, Berlin 2009. Das Virtuelle Skriptorium St. Matthias. In: Libri pretiosi 14 (forthcoming).

Beyond TEI: Returning the text to the reader

Wittern, Christian

Much research and practical effort has gone into the development and maintenance of a digital format that could form a stable foundation for texts in the digital age; the results of this work in the form of the Guidelines for Electronic Text Encoding and Interchange have been widely adapted in the community.

While this does can indeed serve as a foundation for a digital edition of a text, the publication of texts encoded in such a way is still much less well understood and researched. The most common practice for digital publication today is to either publish in some form of

web accessible form, with CD-ROM publication quickly becoming obsolete. In some rare cases, the XML source form of the edition is also available.

For a researcher, this situation is in some respects much worse than it was when critical editions were published only in print, since in most cases the texts can only be browsed online (and every site has it's own idiosyncratic way of displaying and navigating a text) and not physically owned. This not only invalidates many of the potential advantages of digital texts, namely, making the digital edition available for machine mediated analysis, but even denies the reader the most basic form of scholarly activity, that is "active reading" or annotation of the text. What is needed here is the digital equivalent of a "college edition" of a text (and yes, we need to and can do much better than simply converting the text into ePub for consumption by electronic reading devices, but nevertheless this option also deserves attention).

To remedy this situation, a new publication form for digital texts is proposed. At the core, this is a plain text format that only contains very few traces of markup, but serves to make the textual content available to the reader. The text is published through a distributed version control system, which allows the researcher to create branches, annotate, edit or translate the text without losing the connection to the established digital edition and thus to all the other researchers, that are working on this text. If there are differing editions of a text, these editions can be represented as 'branches' in this system, but the assumption is still that there is one privileged 'master' branch that corresponds to a reading text in a critical edition.

In some respects, this text similar to publishing a college or paperback edition of the text established in a critical edition: The user knows that the text is based on a rigorous editing process and thus forms a safe foundation for further research, but at the same time is not burdened with all the details that might get in between her and the text, but has at every stage of his work the possibility to refer back to the critical edition if that becomes necessary. In some other respect, it resembles more the interactive communities or "social networks", that have sprung up recently and already carry a significant amount of scholarly communication. There is however a critical difference: In the model described here and implemented as a proof of concept as part of the Mandoku project (http://www.mandoku.org) the researcher, who publishes annotations in form of additional 'branches' of the master branch of a text retains control and ownership of all these additions, which constitute an essential part of his scholarly work, without compromising the ability to quickly share the results with interested colleagues.

Earlier versions of these experiments used the TEI XML format as base for the texts, but it turned out to be a bad fit to the line-oriented model of texts used in version control systems. Currently, an enhanced version of the Emacs org-mode (http://www.orgmode.org) file format is used. This has the additional advantage of providing a flexible user interface as well as options for direct export to popular other text formats, such as HTML, PDF, OpenOffice XML and DocBook XML. A back converter to TEI XML, that will offer the option to roll the different "branches" of texts maintained in the version control system back into one single file is planned.

Bibliography:

- Christian Wittern, "道藏輯要の編纂と電子化をめぐる諸問 題", in: 麥谷邦夫(編)　三教交涉論叢續編、京 都 2011, p.471-499. (Digital Editions of Premodern Chinese Texts: Problems and Solutions – Exemplified using the Dàozàng jíyào, in Japanese)
- Christian Wittern, "Towards building a Digital Tripitaka in East Asia", in: The Millenium Tripitaka Koreana. Rediscover the Value. Gwangjon (Korea), 2010, p.97-134.
- Christian Wittern, "Mandoku – An Incubator for Premodern Chinese Texts – or How to Get the Text We Want: An Inquiry into the Ideal Workflow", in: Digital Humanities 2010. Conference abstracts. London, 2010, p. 271-273.
- Christian Wittern, "Text and Tradition in Chinese Buddhism", in: Research on the Chinese Materials in the Berlin Turfan Collection: Continued, edited by Tsuneki Nishiwaki, Kyoto 2009, p. 105-111.
- Christian Wittern, Arianna Ciula and Conal Tuohy, "The making of TEI P5", in: Literary and Linguistic Computing, Vol. 24 (3) 2009, 281-296.

Poster and Demos

‘amalia : an eSciDoc based solution to manage the production, processing and publishing workflows of our TEI data.

Beaugiraud, Valérie; Gedzelman, Séverine; Ingarao, Maud; Magué, Jean-Philippe; Saïdi, Samantha

At the ENS Lyon, various teams have chosen TEI to represent their data - corpora, critical editions, etc., with the benefits we know in terms of expertise sharing, strength of XML technologies, long term preservation ... However, the use of the standard is not sufficient to avoid a certain complexity when projects are growing : with the issue of file versioning (which can of course always be solved by an ad hoc software) appear issues of dependence between these versions : for instance, when a XSL file has been modified, which data files have to be reprocessed? What are all the other operations to be launched? This complexity is increased by the problems of authentication and authorization, and by the great variety of software tools involved in the production, exploitation, visualization and / or publication of these data. As part of a sharing policy within the ENS, we develop ‘amalia, a solution based on the platform eSciDoc to address this problem of complexity. eSciDoc is a data management system which provides functionalities to organise data, associate metadata to data, manage versioning, and control access. ‘amalia is a web application which elaborates on eSciDoc and allows the definition of data workflows, i.e. chains of transformations and/or actions operating on data. The actual realisation of the actions orchestrated by ‘amalia is not constraint and may rely on web service, local script, … ‘amalia can be conceived as a glue between various subsystems where the data they work on is ultimately managed by eSciDoc. We will propose a demonstration of ‘amalia and a feedback on the challenges and benefits of this solution in our own experience. We will first show the first workflow we defined, which was for the project Hyperdonat. This workflow transform from ODT to XML TEI and from XML TEI to HTML and index the XML TEI in the XML database BaseX. We will also show other workflows at various stages of development, and involving other third-party systems such as Dinah.

Bibliography:

- Loiseau, Gréa & Magué, 2010. Dictionnaires, théorie des graphes et structures lexicales, in Revue de Sémantique et Pragmatique, 27. pp 51-78.
- Heiden, Magué & Pincemin, 2010. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement, in Statistical Analysis of Textual Data - Proceedings of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010.

It's all about Integration and Conceptual Change

Burr, Elisabeth; Kovacs, Pascal; Potapenko, Elena

The project which we would like to present is a project which originally evolved out of three different types of challenge:

- The wish to make a substantial contribution to the 600th anniversary of the University of Leipzig, celebrated in 2009, by analysing the contribution of the 19th century Leipzig school of Young Grammarians to the development of Romance linguistics;
- The need to teach students of traditional Humanities degree courses not only linguistics but also research methods and the writing of academic papers;
- The intention to integrate Digital Humanities components in traditional modules of Romance linguistics and to foster conceptual change.

The project was carried out together with students of Romance Linguistics in collaboration with a student of Business Information Systems and an external designer. The portal which has been created in the framework of this project is supposed to be not only a means of presenting the project results, but also to function as a teaching tool and as a test bed for further developments. Although the project is ongoing, the portal which is available at http://www.culingtec.uni-leipzig.de/JuLeipzigRo/ can be used as it stands. The portal through its four levels tries to reflect the process of scientific research and the writing of an academic paper. As all four levels as well as the data they contain are interlinked, the portal mirrors also the circularity which is characteristic of academic paper writing in the digital age. The poster will explain the four levels which make up the portal (sources, extract from sources, ontology, texts). Particular attention will be given to the Digital Humanities components, i.e. the TEI markup of primary texts and the bibliographical database, as their integration in traditional modules of Romance linguistics foster the conceptual change which is needed if computer technologies are to be exploited meaningfully when doing research and writing an academic paper.

Bibliography:

- Burr, Elisabeth (ed.) (2011): Sociolinguistique urbaine - Identités et mise en mots [together with Sabine Bastian and Thierry Bulot]. München: Meidenbauer.
- Burr, Elisabeth (ed.) (2009): Sociolinguistique urbaine et développement durable urbain. Enjeux et pratiques dans les sociétés francophones et non francophones [together with Sabine Bastian and Thierry Bulot]. München: Meidenbauer.
- Burr, Elisabeth (ed.) (2009): Tradizione & Innovazione. Dall'italiano, lingua storica e funzionale, alle altre lingue (= Quaderni della Rassegna 59). Firenze: Cesati.
- Burr, Elisabeth (ed.) (2008): Mehrsprachigkeit in frankophonen Räumen / Multilinguisme dans les espaces francophones [together with Sabine Bastian]. München: Meidenbauer.
- Burr, Elisabeth (2005): "Multilingualism and the Web", in: Nicolás, Carlota / Moneglia, Massimo (eds.): La gestione unitaria dell'eredità culturale multilingue europea e la sua diffusione in rete. Atti della conferenza internazionale CLiP 2003 Computers, Literature and Philology , Firenze 04.-06.12.2003. Firenze: Firenze University Press 201-214.
- Burr, Elisabeth (ed.) (2005): Tradizione & Innovazione. Il parlato: teoria - corpora - linguistica dei corpora. Atti del VI Convegno Internazionale della SILFI, Duisburg 28.06.-02.07.2000 (= Quaderni della Rassegna 43). Firenze: Cesati.
- Burr, Elisabeth (in print): "Planification linguistique et Féminisation", in: Baider, Fabienne / Elmiger, Daniel (eds.): Intersexion. Langues romanes, Langue et genre. München: Lincom Europa.
- Burr, Elisabeth (ed.) (in print): Tradizione & Innovazione. Integrando il digitale, l'analogo, il filologico, lo storico ed il sociale. Atti del VI Convegno Internazionale della SILFI, Duisburg 28.06.-02.07.2000 (= Quaderni della Rassegna). Firenze: Cesati.

Wandering Jew's Chronicle

Cummings, James

This poster will describe and demonstrate work done in creation of an online archive for research into the Wandering Jew's Chronicle (WJC). The WJC is a printed ballad published between 1634 and circa 1820 which survives in 22 known copies of 15 editions. These are held in ten libraries in Britain and the USA. The ballad itself outlines the succession to the throne of England from William I to a variable contemporary monarch depending on its date of publication. More specifically these are from the reign of Charles I until that of George IV, taking in seven monarchs in continuations from a core text. The succession of these monarchs is narrated by the supposedly immortal Wandering Jew of European legend. There is immense scholarly interest not only in the subject matter, but textually in the pattern of variations, the length and breadth of its publication and distribution. For a digital humanities perspective the textual history and relationships pose interesting problems for collation and textual analysis. Each one of the editions inherits a basic core text: some of these editions incorporate common continuations or variations, while others are textually idiosyncratic. The editions of the WJC are not only textually but also graphically interesting as most of the editions are illustrated with woodcuts of the monarchs described, and while some editions share woodcuts in common others employ copies or individual illustrations. The poster and demonstration will introduce the benefits of having gathered all the material relating to the WJC in a single place while demonstrating the technologies used to create the research archive. Surviving copies of the WJC are scattered: variously held in the Bodleian Library; the British Library; Cambridge University Library; Magdalen College Cambridge, Pepys Library; The University of Texas at Austin Harry Ransom Library; and the Brown University Library. The WJC research archive has created a digital archive in which surviving editions are united under a single authoritative citation and represented by:

- archival quality images
- transcriptions marked up in TEI P5 XML
- tools for comparing variations between texts
- bibliographical metadata
- scholarly commentary.

It is hoped that by providing all of this in one location research into WJC can flourish in new and interesting ways. This resource helps to foster digital humanities research through tracing and expressing bibliographical, textual and iconological relations across a corpus of copies, variant editions, and versions of ballad texts, including their images and tunes. It is a valuable resource for those researching textual genealogy in the earlymodern period. It will impact the research of scholars of folklore, balladry, historiography, book history and textual studies.

Bibliography:

- Cummings, J.C., 2011, "The materiality of markup and the Text Encoding Initiative", in Digitizing Medieval and Early Modern Material Culture: Text ed. Brent Nelson and Melissa

Terras, New Technologies in Medieval and Renaissance Studies Series, Medieval and Renaissance Texts and Studies. [Forthcoming]

- Cummings, J.C. and Mittelbach, A., 2011, “The Holinshed Project: Comparing and linking two editions of Holinshed’s Chronicles”, International Journal of Humanities and Arts Computing, DHRA 2009 Proceedings
- Cummings, J.C., Rahtz, S.P.Q., McKnight, J., Platinský, L., Werla, M., Stanisławczyk, M., and Parkoła, T., 2010, "OxGarage: An experimental web and RESTful document conversion service" Poster for the 2010 Conference and Members’ Meeting of the Text Encoding Initiative Consortium
- Cummings, J.C., 2010, "The TEI NewsFeed System", Poster for the 2010 Conference and Members’ Meeting of the Text Encoding Initiative Consortium
- Cummings, J.C., Parkoła, T., Stanisławczyk, M., Werla, M., and Psohlavec, T., 2009-12-01, "Report on the Documentation and Use of the ENRICH Garage Engine and on Best Practice in handling of Unicode and Non-Unicode Data", ENRICH Project Deliverables, D3.4. See http://snipurl.com/enrich-d3.4
- Cummings, J.C., and Mittelbach, A., 2009-09-07, "The Holinshed Project: Comparing and linking two editions of Holinshed's Chronicles", Digital Resources in the Humanities and Arts 2009, Queen's University Belfast, http://ora.ouls.ox.ac.uk/objects/uuid%3A223613fd-7481-4fe6-a218-abdb4287b960
- Cummings, J.C., 2009-06-05, "Converting Saint Paul: A new TEI P5 edition of The Conversion of Saint Paul using stand-off methodology", Literary and Linguistic Computing 24:3 307-317 http://llc.oxfordjournals.org/cgi/content/abstract/fqp019
- Cummings, J.C., 2009-04-29, "The William Godwin’s Diaries Project", Jahrbuch für Computerphilologie 10 (2008) http://computerphilologie.de/jg08/cummings.pdf

Patent policies in Dingler's »Polytechnisches Journal« - Exemplary tagging of names, dates and places

Hug, Marius; Gödel, Martina

This poster presentation aims to introduce our digitization project »Dingler-Online« to the TEI community. The project is located at the Institut für Kulturwissenschaft, a department of the Humboldt-Universität zu Berlin and sets out to digitize »Dingler’s Polytechnisches Journal«, which was originally published between 1820 and 1931. Aside from the digitization of the journal’s images, we encode the OCRed text according to the TEI guidelines TEI-P5. At the moment (April 2011) 219 books – which is around 220 million characters – are freely available on the web via www.polytechnischesjournal.de/. Due to its technical background a very important element of the journal are patent applications. At this time (April 2011) our patentlists cover more than 24.000 entries. Hence we are dealing with a great collection of data-sets, which are both compact and expansive in the sense, that each single entry is rather short, but still contains persons, titles, places and dates, that again may point to other sources. Therefore apart from the patentlists’ importance for research activities, these lists are indeed a great challenge for thorough TEI-tagging. This refers both to conventions for our service provider and to our semi-automatical encoding in which we make extensive use of xpath queries and regular expressions. Furthermore our encoding of patentlists is the basis for different visualization approaches, which aim to make our edition a user-friendly platform inspiring a broad use not at all restricted to researchers but open to an interested public in general.

Bibliography:

- Hug, Marius: Auf der Suche nach dem Patent – Ein Blick auf J. G. Dinglers »Polytechnisches Journal«. In: Albert Kümmel-Schnur, Christian Kassung (Hg.): Bildtelegraphie. Eine Mediengeschichte in Patenten (1840-1930). vorauss. Juli 2011.
- Hug, M./Wiegand, F./Kassung, Ch.: Wer kommt in den 1820er Jahren auf die Idee, fliegende Eisenbahnen zu bauen? – Patentanmelder im Polytechnischen Journal. Projektvorstellung auf dem Workshop: Personen – Daten – Repositorien. Berlin-Brandenburgische Akademie der Wissenschaften, 29. September 2010.
- Hug, Marius/Kassung, Christian: A Matter of Communication in the 19th Century. Poster presentation at »Digital Humanities 2010«. London, 7.–10. Juli 2010.
- Hug, Marius/Kassung, Christian/Meyer, Sebastian: Dingler-Online – The Digitized »Polytechnisches Journal« on Goobi Digitization Suite. In: Digital Humanities DH2010. Conference Abstracts. London, 2010, S. 311–313.

Editing Opera: Challenges of an Integrated Digital Presentation of Music and Text based on "Edirom" and TEI (OPERA – Spektrum des europäischen Musiktheaters, Universität Bayreuth / Edirom – Digitale Musikedition, Universität Paderborn)

Münzmay, Andreas; Daniel, Röwenstrunk; Droese, Janine; Seuffert, Janette

Opera editing so far has been undertaken either under a musicological point of view – focussing on the musical score – or from the literary scholar's standpoint – considering the libretto as a literary text genre. This 'splitting' is the result of the specific structural organization and historical transmission of each of the two textual systems. An eighteenth century libretto, for instance, is structured by scenes and verse lines, while the musical score of the same piece is organized by musical numbers and bars. In addition to this, the 'literary' and 'musical' text traditions of one and the same work of the lyric stage often are anything but congruent, which is, of course, due to performance and adaptation practices, but also to the fundamental differences between the two text types. On the other hand, it is evident that an item of music theater – a vast generic field which includes not only 'normal' opera composed from one end to the other, but also more 'complicated' genres as e. g. for comic opera, Singspiel, operetta, drama with incidental music, melodrama, or ballet – comes down to the editor not only through musical sources, but by a (mostly large) bundle of literary and musical ones. OPERA aims to handle the diversity of the source types in a way that on the one hand includes all relevant source types and considers them of equal value, and on the other hand avoids mutual contamination. OPERA editions are presented in the form of hybrid editions with the scores being published in printed volumes, while the text editions and critical reports are in an electronic form visualizing the sources and establishing a complete interlinking of the editions with the sources and critical reports. For this purpose, OPERA uses the music edition software Edirom, which has been developed by the Edirom project at Paderborn University and recently was essentially expanded in regard to the incorporation of text editions. The poster and tool presentation gives an insight into two editions OPERA currently works on: the Italian opera Prima la musica e poi le parole (Vienna 1786) by Antonio Salieri (music) and Giambattista Casti (text), edited by Thomas Betzwieser and Adrian La Salvia, respectively, and the French opéra comique Annette et Lubin (Paris

1762) by Justine Favart (text) and Adolphe Blaise (music), with the text and music edition being prepared by Andreas Münzmay. These examples highlight the editorial, programming and encoding strategies ensuring the integration and interlinking of TEI-based text edition, score edition, source images, and critical apparatus. Among the features that have been developed to this end is a navigation tool which allows to call up whatever bar of the score or whatever verse line of the libretto and visualize synchronously any corresponding textual and musical object. Likewise, one common critical apparatus comments on both the text and music edition, and is accessible from either of the two directions. For this purpose, it is necessary in turn to incorporate musical structure information such as bar numbers, and link information with respect to the source images and critical apparatus entries into the TEI-based text edition section of an edition's XML source code.

Bibliography:

- Thomas Betzwieser: "Editing musical quotations: the paradigm of Antonio Salieri's *Prima la musica e poi le parole (*1786)", in: Philomusica on-line 9/2 (Atti del VI Seminario Internazionale di Filologia Musicale "La filologia musicale oggi: il retaggio storico e le nuove prospettive"), S. 245-259 (http://riviste.paviauniversitypress.it/index.php/phi/article/view/919)
- Die Tonkunst 5 (2011) No. 3 (July 2011), Themenheft: Perspektiven Digitaler Musikedition.
- Peter Stadler und Joachim Veit (eds.): Digitale Edition zwischen Experiment und Standardisierung. Musik - Text - Codierung, Tübingen: Niemeyer, 2009 (= Beihefte zu editio, Bd. 31).

TXSTEP - an integrated XML-based scheme for scholarly text data processing

Ott, Wilhelm; Ott, Tobias

TXSTEP offers an interactive XML-based interface to the proven and powerful routines of TUSTEP, the Tübingen System of Text Processing programs. For more than 35 years TUSTEP is being developed and maintained at Tübingen University's computing centre. TUSTEP is a scripting language as well as a publishing system for the humanities, up until today unmatched in its overall performance and flexibility. TUSTEP primarily addresses users in the fields of the textprocessing humanities, such as computerlinguists, -philologists and editors. For more information, see www.tustep.org. But, since it's genuine syntax is proprietary, not intuitive and supposed tobe difficult to learn, users tend to help themselves with other - often less effective - tools or less specific programming languages. TXSTEP now gives a good answer to this situation by providing a user-friendly XML-syntax, allowing beginners and advanced programmers to use the whole scope of TUSTEP services in a modern, established, programmers environment. The benefits are obvious: support of an open standard, widespread dissemination, programming in every other XML-editor, syntax highlighting, code completion and intelligible APIs. Moreover, TXSTEP is aided by the fact that there is no need to change the program's actual core. TUSTEP itself is open source as TXSTEP is going to be as well. The purpose of TXSTEP, as well as of TUSTEP, is not to provide ready-made solutions for pre-defined problems. It "only" provides program modules for the basic functions of text analysis and processing. It is the user who has to combine them in order to

obtain the solution to a problem at hand. This is the prerequisite that he can take over the responsibility for every detail of the results obtained by computer application. One of the features of TXSTEP is its capability to process almost all forms of textual data, whether this being XML-data or plain text files. Wherever there is textual data that has to be processed in the first place in order to gain TEI-data or to enhance the markup of insufficiently tagged XML data, TXSTEP is at its place. The proposed demo is based on a prototype and shows the achieved state of our work in progress. It will demonstrate TXSTEPs functionality on the basis of tasks which cannot easily be performed by existing XML tools, including problems presented recently on the TEI list.

Bibliography:

- Digital Publishing: tools and products in: Poiesis & Praxis: Internat. Journal of Technology Assessment and Ethics of Science Vol 5 Nr. 2 (2008), S.81-112.

Virtual Scriptorium St. Matthias

Vanscheidt, Philipp; Scholzen, Sabine

The project Virtual Scriptorium St. Matthias intends to reunite the worldwide scattered codices from the library of the Benedictine abbey St. Eucharius or St. Matthias in Trier electronically. The project is realized at Stadtbibliothek and Stadtarchiv Trier as well as at the Center for Digital Humanities at the University Trier since summer 2010. About 450 codices from the period between the eights and fifteenth century will be digitized in three years. These codices concern a wide range of topics from various traditions. Beyond theological and religious writings you find a large amount of latin classics like Cicero, Priscian, Sallust or Martianus Capella. A prestigious example for the inculturation of ancient and pagan spirit is an illustrated edition of Aesops fables. No other abbey possessed as many manuscripts of Hildegard of Bingen as St. Matthias. You find also three important specimina of Dectretum Gratiani. One of them includes 60% of all glosses ever written on this work. But the richly illustrated Trierer Apokalypse from carolingian times maybe the most famous of all these codices. The project Virtual Scriptorium St. Matthias will present an electronic catalogue that sums up the knowledge from older descriptions and combines them with a presentation of the digitized codices. In this context TEI is used as a standard of XML description of manuscripts. The amount of objects requires a synchronization of these descriptions with a dynamic database to correlate them with other digitized catalogues, editions and databases like the PND. The results will be integrated in Manuscripta Mediaevalia and TextGrid. In this way the project will not only provide images and metadata but will also be included into a virtual working space in where further research and exploring will be possible, e.g. with TEI concurrent transcriptions of selected works.The project homepage www.stmatthias.uni-trier.de will be released on the first of August 2011 on a trial base. The project should be presented in a short talk and a poster. The poster will cover the project thoroughly while the short talk is supposed to sketch the advantages and some practical limits of TEI in such an enterprise.

Bibliography:

- Geschichte in Metaphern, Berlin 2009. Das Virtuelle Skriptorium St. Matthias. In: Libri pretiosi 14 (forthcoming).

Quality Assurance of Large TEI Corpora

Wiegand, Frank; Geyken, Alexander

The DFG-funded project Deutsches Textarchiv (DTA) started in 2007 and is located at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). Its goal is to digitize a large cross-section of German texts from 1650 to 1900. The DTA presents almost exclusively the rst editions of the respective works. Currently there are more than 700 texts transcribed, most of them by non-native speakers using the double keying method. Even though our corpus of historic text exhibits very good quality, enough errors still occur in the transcription, in the markup, or even on the level of presentation. Due to the heterogeneity kind of the corpus, e. g. in terms of text sorts -- novels, prose, scientific essays, linguistic reference works, cook books &c. -- there is a strong demand for a collaborative, easy to use quality assurance environment. Our poster and tool demonstration will provide an insight into the DTA QA workflow.

Best Practices for TEI in Libraries

Kevin Hawkins, Michelle Dalmau, Melanie Schlosser

The TEI Consortium's Special Interest Group (SIG) on Libraries (www.tei-c.org/Activities/SIG/Libraries/) has recently completed a major revision to Best Practices for TEI in Libraries (purl.oclc.org/NET/teiinlibraries). The revised Best Practices are stored in ODD files and contain updated versions of the widely adopted encoding "levels", which span from fully automated reformatting of print content to rich encoding to support content analysis and scholarly uses. They also contain a substantially revised section on the TEI header to support greater interoperability between text collections and MARC records. Schemas for each encoding level, derived from the appropriate ODD, provide a mechanism to better ensure conformance and interoperability of digital text. Principle contributors to this document will present a poster summarizing the encoding levels in the Best Practices, how the schemas work in relation to the TEI Guidelines, and what digitization workflows are envisioned for use with them.

The project "Berlin intellectuals 1800-1830" between research and teaching

Baillot, Anne; Seifert, Sabine

The project "Berlin intellectuals 1800-1830" is a DFG-funded 5-year project based on unedited manuscripts (mainly, but not exclusively letters). It aims at gaining an in-depth insight into intellectual networks in Berlin at the beginning of the 19th century. The corpus we focus on allows to identify intellectual affinities and their evolution according to the

scientific or political context of the moment: to that extent, history of ideas plays a central role in the project. Yet, its main focus lays in literary history. We aim at describing communication strategies and mechanisms as exactly as possible, working on the very borderline between private and public texts. Letters play a central role, as they are themselves literary objects that, especially in the context of Germany in the early 19th century, have an ambiguous position in the literary field. Many of them have been preserved in order to be published, some of them have indeed been partially published, often even partially re-written. By having access to the original manuscript, we are now being given the chance to re-write literary history: demonstrate the biases of earlier, faulty editions and give a larger public access to first-hand documents. The workflow in the project consists of defining relevant areas of research, searching for material, selecting the most interesting, digitizing as well as transcribing it and - then what?

Confronted with this question, we moved away from a traditional paper-edition and, in a second step, from a basic online-edition to develop a TEI-based concept that fulfills the varying demands of our text corpus. After dealing closely with the TEI Guidelines and discussing what data and metadata we want to encode, we designed our own XSL stylesheet as well as TEI-schema that concentrates on the requirements of encoding correspondence and has a letter-specific manuscript description. Using the Oxygen XML Editor, the TextGridLab and integrating databases proved to be convenient to our needs. Although now being experienced in encoding manuscripts and letters according to the TEI, we encounter some aspects that still remain problematic.

During the course of the project, not only project members have been involved but also a large number of graduate philology students. In the phase of development of the TEI-schema and conception of a frontend for the intended online-edition, several documents relevant to the project were offered as assignments to students in German and European Philology at the Humboldt University. The students reflected the hermeneutical dimension of the editorial process with a remarkable maturity, showing their ability to implement the theoretical discourse they have been made familiar with in their Humanities studies in the digital medium. This experience tends to suggest that even traditional curricula are nowadays in the position to include a digital part.

In this poster, we present the whole editorial process from the archival discovery to the publication of the material, from defining a TEI-schema that meets the specific and sometimes complex requirements of encoding correspondence to the final creation of the frontend. We also give a record of the pedagogical choices that were made to make the student's input fruitful. The tensions of the discussions, between technicalities and hermeneutics, are being analyzed. We finally make some suggestions regarding the sustainability of digital humanities in European humanities curricula.

Bibliography:

- Friedrich der Grosse – Oeuvres philosophiques/Philosophische Schriften (Hg.), bearb. zs. mit B. Wehinger, Bd VI von: Friedrich der Grosse – Potsdamer Ausgabe Französisch/Deutsch, hg. v. A. Baillot, G. Lottes u. B. Wehinger, Akademie-Verlag, Berlin, 2007, 524 Seiten

- „Wie rehabilitiert man einen Schriftsteller und wozu? Das Beispiel unerschlossener Briefwechsel aus dem Umfeld des Dichters Ludwig Tieck, des Philosophen Karl Solger und des Historikers Friedrich von Raumer“, in: Dokument/Monument. Textvarianz in den verschiedenen Disziplinen der europäischen Germanistik. Akten des 38. Kongresses des französischen Hochschulgermanistikverbandes A.G.E.S., hg v. F. Lartillot u. A. Gellhaus, Peter Lang Verlag, Bern/etc. 2007, S. 103-126

William Godwin's Diary

Cummings, James

William Godwin (1756-1836) wrote a diary that consists of 32 octavo notebooks. The first entry is for 6 April 1788 and the final entry is for 26 March 1836, shortly before he died. The diary is a resource of immense importance to researchers of history, politics, literature, and women’s studies. It maps the radical intellectual and political life of the late eighteenth and early nineteenth centuries, as well as providing extensive evidence on publishing relations, conversational coteries, artistic circles and theatrical production over the same period. One can also trace the developing relationships of one of the most important families in British literature, Godwin’s own, which included his wife Mary Wollstonecraft (1759-1797), their daughter Mary Shelley (1797-1851) and his son-in-law Percy Bysshe Shelley (1792-1822). Many of the most important figures in British cultural history feature in its pages, including Anna Barbauld, Samuel Taylor Coleridge, Charles James Fox, William Hazlitt, Thomas Holcroft, Elizabeth Inchbald, Charles and Mary Lamb, Mary Robinson, Richard Brinsley Sheridan, William Wordsworth, and many others.

The William Godwin's Diary website <http://godwindiary.bodleian.ox.ac.uk/> presents richly marked up TEI P5 XML texts via Apache Cocoon and the eXist XML database with custom Xqueries and XSLT. In addition to the transcribe texts and high resolution zoomable images, the site includes many extracted data tables based on the underlying markup. All the materials on the website, including the underlying TEI P5 XML, are available under a Creative Commons Attribution Non-Commercial license. This poster will present the underlying architecture of the project, the compromises made and difficulties encountered. The poster will be accompanied by a demonstration of the website and show how the site works underneath for those interested in those aspects.

Bibliography:

- Cummings, J.C., 2009-04-29, "The William Godwin’s Diaries Project",Jahrbuch für Computerphilologie 10 (2008) http://computerphilologie.de/jg08/cummings.pdf

Workshops and Tutorials

Analysing Electronic Dictionaries with TEI

Dietmar Seipel

The workshop deals with developing tools for analysing morphological processes in word formation that might be applicable for corpus research of complex words. Topics:

- tools for morphological analyses
- lexicographical structures
- word morphology
- language change
- variance between and within dictionaries
- modelling morpheme structures
- construction and analysis of meta-dictionaries in TEI
- parsing techniques for information extraction
- analysis of morpheme networks
- declarative languages for analysis

Tightening the representation of lexical data, a TEI perspective

Laurent Romary

The “Print Dictionary” chapter of the TEI guidelines was initially conceived to represent human readable dictionaries, in particular in the case of the digitization of existing paper dictionaries. Its neat hierarchy of lexicographic concepts has also made it a very practical framework for representing machine-readable lexAleical data, even in very specific NLP applications. Still, as is often the case with the TEI guidelines in general, there are often many different ways to represent the same phenomenon, which is a potential hindrance for ensuring full interoperability between TEI based lexical applications. Finally, the recent publication of ISO standard 24613 (LMF – Lexical Markup Framework) has shown the necessity to identify a TEI based serialisation that would optimally map the concepts of the LMF meta-model.

The workshop will bring together representatives from various lexical database (both human- and machine-readable ones) initiatives to compare their practices in implementing the TEI guidelines. The format will be based on short presentations focusing on specific lexical phenomena and suggesting precise guidelines for their representations. The expected output of the workshop is a set of application guidelines to be used for machine readable dictionaries.

Contributions are expected in the wide variety of potential lexical structures. The following lists should just be seen as mere suggestions for possible topics:

- Representation of morphological information
- Additional grammatical features
- Constraints and dependencies between features

- Implementation of LMF components
- Standardisation of data categories (link to ISOCat)
- Tightening sense representation
- Multilingual lexical, link structures
- Derivations and related entries

Combining Music Notation and Text – Encoding and Rendering MEI in TEI

Johannes Kepper

A large number of TEI projects deal with texts containing music notation. These inclusions range from short melody or rhythm snippets to several pages of music. During the last year, the SIG Music has worked on the integration of TEI and MEI, requested a new element in the TEI and produced some guidelines. The proposed workshop has two goals.

First, the current model will be discussed and evaluated against different kinds of source documents. It is intended to have an open discussion with participants currently not involved in the working group and work on their documents as well. The group has already had the opportunity to collect examples and opinions from several scholars; however, this workshop will incourage new contributions and will introduce the proposed model to the public.

The second part of the workshop will focus on rendering MEI. Currently, there is only a tool for converting MEI-encoded mensural notation into SVG, but nothing similar for Common Western Music Notation, which is a serious issue in TEI with MEI based workflows. An international group of specialists has already started to work on this. The problem of rendering MEI into SVG is indeed closely related to the work done by the SIG so far; many examples of music notation in text offer a wide range of complexity. Working on these can make a contribution to more generic stylesheets and tools. Although this second part of the workshop may be less suited for beginners, it is intended to be open to the public in order to gather diverse requirements from projects as varied as possible.

Using the Scalable Architecture for Digital Editions (SADE) for the digital presentation of TEI encoded texts

Alexander Czmiel

The Telota working group of the Berlin-Brandenburg Academy of Sciences and Humanities wants to offer a tutorial on the Scalable Architecture for Digital Editions (SADE). The tutorial aims to introduce SADE as publication framework for any kind of text based digital resources. It will show the participants the various deployment methods and explain how the different components (XML database, image viewer, front-end etc.) of SADE work together. In the hands-on-section of the tutorial the participants will learn how to deploy SADE themselves and how to adjust the provided XQuery, XSLT and CSS-scripts for their own needs. As SADE comes with native support for TEI encoded texts it serves as an appropriate tool to query, transform and publish these texts. The included image viewer digilib in turn is a proper tool for text-image-linked presentations.

The target audience are (digital) humanists with good knowledge of XML/TEI and HTML and at least basic knowledge of XSLT, XQuery and XPath, who plan to produce an own digital edition based on the TEI guidelines.

References:

- http://www.bbaw.de
- http://www.bbaw.de/telota/
- http://www.bbaw.de/telota/sade
- http://digilib.berlios.de/

Tuning oXygen XML Editor for TEI

George Bina

This tutorial will provide an overview presentation of the oXygen features going into details on the TEI specific support and on the parts that are generally used in relation with TEI work. oXygen comes with built-in support for TEI but this support is not hard coded, it is just a default configuration that can be extended/modified/customized according to your needs. I will explain what the default TEI support consists of and how this can be customized and shared within a group of users. We will discuss also validation technologies like Schematron or NVDL, working with XML databases, processing/querying documents, etc. I am open also to any topic from the audience and part of the tutorial will be an open discussion on any TEI - XML - oXygen related topic.

Using TILE to build links between text and images in TEI projects

Dot Porter

Aimed at student and faculty scholars with intermediate to advanced experience using TEI or XML, this course will provide an introduction to using TILE, the Text-Image Linking Environment, to build various types of links for TEI projects (http://tileproject.org/). Working with the Image markup plug-in and the Semi-automated line recognizer in the online installation of TILE, during the morning we will look at a variety of different TEI files that represent links between text and image, for example manuscripts and transcriptions of the texts contained in them, paintings and poems written to describe them, and illustrations and annotations describing them. During the afternoon students will work with their own TEI files in TILE. The instructor will provide materials for students who do not bring their own.

An introduction to working with the TextGridLab

Oliver Schmid, Celia Krause and Philipp Vanscheid

The joint project TextGrid aims to support access to and exchange of data in the arts and humanities by means of modern information technology. TextGrid serves as a virtual research environment for philologists, linguists, musicologists and art historians. As a single point of entry to the virtual research environment, TextGridLab(oratory) provides integrated access to specialized tools, services and content. TextGridRep(ository) is a long-term archive

for research data in the humanities embedded in a grid infrastructure, which will ensure availability and access to its research data as well as interoperability.

The workshop will introduce the basics of working with the TextGridLab. A brief theoretic introduction will lay the groundwork for the interactive main section of the course. First of all, the basic knowledge of the project and the infrastructure of TextGrid will be imparted, then the general tools are presented and tested by the participants during the interactive section. Subsequently, the expert tools are brought into focus, prioritizing the work with TEI encoded texts and image files using the Text Image Link Editor. Furthermore, the functional range of the XML Editor and the Metadata Editor is presented as well as further expert tools.

References:

- www.textgrid.de/en/startseite.html (project page)
- www.textgrid.de/en/1-0.html (overview of TextGridLab tools)
- www.textgrid.de/en/1-0/download.html (software download)

Preparing a Critical Edition of an Incunabula with TEI

Guenther Goerz, Josef Schneeberger, Klaus Thoden

Digital editions of incunabula --- books printed before 1500 --- provide a number of challenges. Many practices from the manuscript culture survived in typesetting for some time, e.g. the use of a large number of special glyphs for ligatures and abbreviations. Word breaks at the end of lines are often unsystematic and in many cases there is no hyphenation sign at all. In many cases, spacing between words is rather narrow and sometimes also missing. Furthermore, punctuation marks are missing or put in places where one would not expect them, such that sentence borders cannot always be recognized clearly. Of course, there is no normalized spelling such that many variants of words occur, sometimes even with erroneous spelling.

In this workshop, we will take up a practical example to discuss the decisions to be made to solve the aforementioned problems in the framework of TEI. Our example is the “Deutsche Ptolemaeus”, a geographical work written in early modern German and printed around 1490, of which only two copies are extant. TEI offers facilities for semantic annotations, e.g. for person and place names (named entities), and technical terms. Furthermore, it provides means for linking with lexical resources, glossaries, and comments. Most tools of computational linguistics are only to a certain extent suitable to deal with the particular linguistic variety which is a dialect variant of early modern southern German. Hence, some specific heuristics have to be developed to at least partially automate the tagging process. Furthermore, we will present specific scripts in XSLT to produce different versions of the transcription, XHTML, raw text, and printed editions.