JADH 2018

"Leveraging Open Data"

September 9-11, 2018
Hitotsubashi Hall, Tokyo

https://conf2018.jadh.org

Proceedings of the 8th Conference of Japanese Association for Digital Humanities

TEI 2018

"TEI as a Global Language"

September 9-13, 2018
Hitotsubashi Hall, Tokyo

https://tei2018.dhii.asia

Book of Abstracts
The 18th Annual TEI Conference and Members' Meeting

JADH and TEI Joint Keynote Session

The NIJL Database of Pre-modern Japanese Works

Robert Campbell[1]

Abstract
NIJL (the National Insitute of Japanese Literature) is currently engaged in digitizing, tagging and developing new ways to search the uniquely rich heritage of pre-modern (prior to 1868) Japanese literary documents. In my talk I will aim to introduce this ongoing project, and through doing so will attempt to suggest several directions in which cross-disciplinary collaboration may lead to a deepening of cultural understanding and innovation.

Biography
Director, National Institute of Japanese Literature. Professor Emeritus, University of Tokyo. PhD., Harvard University.

Related URL
https://www.nijl.ac.jp/

[1] National Institute of Japanese Literature

Amsterdam 4D: Navigating the History of Urban Creativity through Space and Time

Julia Noordegraaf[1]

Abstract

A city's cultural industries are a major factor in its economic durability and the wellbeing of its inhabitants. Successful urban agglomerations, such as the city of Amsterdam, are marked by the concentration of artists, performers, and knowledge workers – cultural production contributes to a city's overall capacity for innovation and competition. Scholars in the humanities and social sciences have begun to explore this dynamic. Thus far, however, they have struggled to explain the correlation between the micro level of cultural interaction and the macro level of a Creative City's economic and social success. Many factors complicate any urban setting's spatial and historical fabric. Where are the creative entrepreneurs located? How do they communicate, interact, collaborate, and compete? How do their products find their ways to the consumers? And how do they turn the city into a magnet for other innovators? In the research program Creative Amsterdam: An E-Humanities Perspective, located at the University of Amsterdam's Centre for Cultural Heritage and Identity, scholars from the humanities and computer science collaborate in using data on the various cultural sectors of Amsterdam with digital methods to investigate how cultural industries have shaped Amsterdam's unique position in a European and global context, from the 17th century until the present day. A central focus of the CREATE program is the building of the Amsterdam Time Machine (ATM): a hub for navigating linked historical data on Amsterdam. This web of information on people, places, relationships, events, and objects will unfold in time and space through geographical and 3D representations. In this "Google Earth for the past", users can go back and forth between the city as a whole, specific neighborhoods, streets, or houses, and even zoom in on the pictures that adorned the walls of for instance merchants and regents. The systematic linkage of datasets from heterogeneous sources allows users to ask new questions on, for instance, cultural events, everyday life, social relations, or the use of public space in the city of Amsterdam. As such, it allows scholars to connect specific objects, persons or places to the level of broader social processes in the city as a whole – functioning as a microscope and telescope in one. Such a research environment, in which space is chosen as the point of view, offers an unprecedented opportunity to investigate the relationship between the physical and social space in relation to how it was organized and experienced over time.

In this lecture I present the design and architecture of the Amsterdam Time Machine, and illustrate its research potential by discussing examples of recent research projects on the history of Amsterdam as a creative city.

Biography

Julia Noordegraaf is professor of Digital Heritage in the department of Media Studies at the University of Amsterdam. She is director of the Amsterdam Centre for Cultural Heritage and Identity (ACHI), one of the university's research priority areas, where she leads the digital humanities research program Creative Amsterdam (CREATE) that studies the history of urban creativity using digital data and methods. Noordegraaf's research focuses on the preservation and reuse of audiovisual and digital heritage. She has published, amongst others, the monograph Strategies of Display (2004/2012) and, as principal editor, Preserving and Exhibiting Media Art (2013) and acts as principal editor of the Cinema Context database on Dutch film culture. She currently leads research projects on the conservation of digital art (in the Horizon 2020 Marie Curie ITN project NACCA) and on the reuse of digital heritage in data-driven historical research (besides CREATE in the Amsterdam Data Science Research project Perspectives on Data Quality and the new,

[1] University of Amsterdam

NWO funded project Virtual Interiors as Interfaces for Big Historical Data Research). She is a former fellow of the Netherlands Institute for Advanced Study in the Humanities and Social Sciences and acts as board member for Media Studies in CLARIAH, the national infrastructure for digital humanities research, funded by the Netherlands Organization for Scientific Research, NWO. Noordegraaf currently coordinates the realization of the Amsterdam Time Machine and participates as Steering Committee member in the European Time Machine project that aims to build a simulator for 5.000 years of European history.

Related URLs
http://www.uva.nl/profiel/n/o/j.j.noordegraaf/j.j.noordegraaf.html
https://twitter.com/jjnoordegraaf
http://achi.uva.nl/
http://www.create.humanities.uva.nl/
http://timemachineproject.eu/

Creating Collections of Social Relevance

Susan Schreibman[1]

Abstract

Digital Humanities, and by extension digital humanists, tend towards a culture of open access, interdisciplinary collaboration, and a maker ethos. These disciplinary values position the digital humanities for high impact reaching beyond disciplinary boundaries into more public fora. One might argue that this public-facing ethos is a natural extension of web-based scholarship. Yet, simply putting resources on the web does not necessarily engage the public or publics they wish to reach. The model still used by the majority of digital humanities projects is that a small team designs and creates the resource, and when it is 'finished' (or at least ready for others to view) it is made public.

This talk will focus on a different model, that of participatory design, which involves the public, that is, anybody who might benefit from our scholarship, in the research process.

Participatory Engagement projects provide us with opportunities to rethink our roles as researchers and as educators, about our obligations to those in society who have not had the same opportunities as we have, and how to build meaningful, socially relevant, digital collections for our own and future generations. This talk will explore the philosophy, mechanics, and value of designing digital scholarship within a participatory model.

Biography

Susan Schreibman is Professor of Digital Humanities and Director of the Centre for Digital Humanities, Maynooth University, Ireland. Dr Schreibman has published and lectured widely in digital humanities and Irish poetic modernism. Her current digital projects include *Letters 1916-1923* and *Contested Memories: The Battle of Mount Street Bridge*. Her publications include *A New Companion to Digital Humanities* (2015), *Thomas MacGreevy: A Critical Reappraisal* (2013), *A Companion to Digital Literary Studies* (2008), and *A Companion to Digital Humanities* (2004). She is the founding Editor of the peer-reviewed *Journal of the Text Encoding Initiative* and is a member of the Board of the National Library of Ireland.

[1] Maynooth University

TEI 2018

"TEI as a Global Language"

September 9-13, 2018
Hitotsubashi Hall, Tokyo

https://tei2018.dhii.asia

Book of Abstracts

The 18th Annual TEI Conference and Members' Meeting

Hosted by:
Center for Evolving Humanities, Graduate School of Humanities and Sociology, The University of Tokyo

Supported by:
JSPS Grant-in-Aid Project (S) "Construction of a New Knowledge Base for Buddhist Studies" (15H05725)
International Institute for Digital Humanities

Sponsored by:
<oXygen/> XML Editor

The 18th Annual TEI Conference and Members' Meeting: Book of Abstracts

Edited by Tensho Miyazaki

PDF Version (Sep 7, 2018)

Published by:

Center for Evolving Humanities, Graduate School of Humanities
and Sociology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
http://21dzk.l.u-tokyo.ac.jp/DHI/

International Institute for Digital Humanities
5-26-4-11F Hongo, Bunkyo-ku, Tokyo
https://www.dhii.jp

Table of Contents

TEI 2018 Committees

Program Committee:

- Susan Schreibman, Chair (Maynooth University, Ireland)
- Masahiro Shimoda, Local Host (The University of Tokyo, Japan)
- Kiyonori Nagasaki, Local Host (International Institute for Digital Humanities, Japan)
- Marcus Bingenheimer (Temple University, USA)
- Elisa Beshero-Bondar (University of Pittsburg, USA)
- Alejandro Bia (Miguel Hernández University, Spain)
- Marjorie Burghart (CNRS, France)
- Vinayak Das Gupta (Shiv Nadar University, India)
- Yuta Hashimoto (National Museum for Japanese History, Japan)
- Nie Hua (Peking University, China)
- Asanobu Kitamoto, OC Chair of JADH (National Institutes of Information, Japan)
- Elena Pierazzo (Universite Grenoble Alpes, France)
- Peter Stadler (Paderborn University, Germany)
- Sarah Stanley (Florida State University, USA)
- Toma Tasovac (Belgrade Center for Digital Humanities, Serbia)
- Magdalena Turska (Oxford University, UK)
- Taizo Yamada (The University of Tokyo, Japan)
- Raffaele Viglianti (University of Maryland, USA)
- Pip Willcox (Oxford University, UK)
- Christian Wittern (Kyoto University, Japan)

Local Committee:

- Masahiro Shimoda, Chair (The University of Tokyo)
- Asanobu Kitamoto (National Institutes of Informatics)
- Kiyonori Nagasaki (International Institute for Digital Humanities)
- Toru Tomabechi (International Institute for Digital Humanities)
- Tensho Miyazaki (International Institute for Digital Humanities)

Time Table

September 9 (Sun), Day 1

9:30-13:30	Workshops
13:30-14:30	Lunch Break
14:30-18:30	Workshops

September 10 (Mon), Day 2

9:00-9:30	JADH and TE Joint Opening Session
9:30-12:00	TEI Council Open Meeting
12:00-13:00	Lunch
13:00-16:15	Opening Plenary (shared with JADH)
16:45-19:00	Poster Session & Demonstrations (shared with JADH)
19:00-	Conference Banquet

September 11 (Tue), Day 3

9:00-10:30	SIG Sessions
10:30-10:45	Coffee Break
10:45-12:15	SIG Sessions
13:45-15:15	Conference Sessions
15:15-15:30	Coffee Break
15:30-17:00	Conference Sessions
17:00-17:15	Coffee Break
17:15-18:45	Conference Sessions and Panel
19:00-	Casual Dinner arranged in groups

September 12 (Wed), Day 4

9:00-10:00	TEI Plenary Meeting
10:30-12:00	TEI Members' Meeting
13:00-	Joint Excursions with JADH (free)

September 13 (Thu), Day 5

9:00-10:30	Conference Sessions and Panel
10:30-10:45	Coffee Break
10:45-12:15	Conference Sessions
12:15-13:15	Lunch Break
13:15-14:45	Conference Sessions and Panel
14:45-14:55	Break (Coffee Room closed)
14:55-16:25	Conference Sessions and Panel
16:30-	Closing (Room A1)

Plenary

Digital Texts in Practice

Christian Wittern[1]

As a student of intellectual, religious and cultural developments in areas of the Chinese cultural sphere, my initial drive in engaging with digital texts 30 years ago was to open up the new possibilities offered by the digital medium for the use of researchers, without losing any of the affordances of a traditional printed edition. This requirement includes use of texts for reading, translating, annotating, quoting, and publishing, thus integrating with the whole of the scholarly work flow.

At that time theories of electronic texts started to appear and the TEI already had begun the endeavour to create a common text model and interchange specification, for the most part based on European languages. For East-Asian texts, things where much more complicated due to different and quickly evolving character encoding standards, different textual traditions and approaches to text editing, as well as different institutional embedding.

In this presentation, I will look back at these developments, firstly to recount some of the history, albeit from a strictly personal perspective, but also to take stock of the situation and consider where we are now, how we got there and what remains to be done to realize the dream of the universal digital text, easily shared and annotated, but still tractable, verifiable and authoritative.

[1] Kyoto University, Japan

[Day 1 (Sun)] Full-day Workshops

Introduction to EpiDoc

Hugh Cayless[1], Yukiko Kawamoto[2], Elli Mylonas[3], Kazuhiro Takeuchi[4]

(Room A2)

EpiDoc (http://www.stoa.org/epidoc/gl/latest/) is one of the most successful TEI customizations in existence. It was originally developed to encode Greek and Roman documents written on stone or other non-perishable materials, but has been generalized to work with other types of document where the scholars who study them care principally about recording not just the transcription, but the text's materiality and the editorial interventions that have been made to establish a readable text. For example, EpiDoc has been used to publish documents on papyrus in a variety of languages, and has been used for other ancient documents in a variety of languages and scripts (including Arabic, Egyptian Hieroglyphs, Hebrew, Mayan Hieroglyphs, and Old Cham).

The wide adoption of Epidoc, as well as the fact that it is a TEI customization, allow its users to take advantage of support and advice from a community of scholars, shared tools for display and analysis, the development of common practices and vocabularies to make it easier to share and re-use Epidoc corpora.

We expect the workshop to appeal both to scholars working on western epigraphy and papyrology in Japan and to those working with similar documents from Japan or other countries with their own epigraphic or papyrological traditions. We also anticipate that there will be an opportunity for participants to discuss their own projects.

The workshop will cover the following topics:

- Brief introduction to XML, TEI and Epidoc
- Basic editing of epigraphic texts.
- OxygenXML software for editing and transforming into HTML for proofreading and display.
- Encoding the history and description of the textual support.
- Advanced Features (apparatus criticus, verse, complex texts). This section will be adjusted based on the interests of the participants.

Participants should have some familiarity with epigraphic, papyrological or similar textual material. An understanding of the Leiden Conventions is helpful but not required. No technical skills are required and scholars of all levels, from students to professors, are welcome.

[1] Duke University, USA
[2] Nagoya University, Japan
[3] Brown University, USA
[4] Osaka University, Japan

TEI Publisher: Power to the Editor

Wolfgang Meier[1], Magdalena Turska[1]
(Room B1)

Abstract
Crossing the divide between encoded XML sources and published digital edition has always been a weak spot for TEI community. TEI Publisher, an eXist-db based application bridges that gap with its implementation of the processing model allowing to create standalone digital editions out of the box.

Publishing an edition so far involved tedious work on complex stylesheets and significant effort in building an application on top of it. Using the TEI Processing Model, customising the appearance of the text is all done in TEI ODD. This approach easily saves thousands of lines of code for media specific stylesheets. eXistdb and its application framework on the other hand take care of all the other core features like browsing, search and navigation.

The proposed workshop intends to introduce the concepts of the TEI Processing Model and provide a tutorial on how to generate custom standalone edition using the App Generator.

Full Description
Crossing the divide between encoded XML sources and tangible, published digital edition has always been a weak spot for TEI community. Recent efforts of the TEI Simple project aimed to bridge that gap with TEI Processing Model idea. TEI Publisher, an eXist-db based application brings the promises of TEI Simple to life with its implementation of the processing model enhanced with an app generator, allowing to create standalone digital editions out of the box.

Publishing an edition from TEI sources so far involved tedious work on complex stylesheets and significant effort in building an application on top of it. Using the TEI Processing Model, customising the appearance of the text is all done in TEI ODD by mapping each TEI element to a limited set of well-defined behaviour functions, e.g. “paragraph” or “heading”. The TEI Processing Model specification includes a standard mapping, which can be tweaked by overwriting selected elements. Rendition styles are transparently translated into different output media types like HTML, XSL-FO, LaTeX, or ePUB. This approach easily saves thousands of lines of code for media specific stylesheets. The power of the eXist-db database and the application framework on the other hand take care of all the other core features like browsing, search and navigation.

The proposed workshop intends to introduce the concepts of the TEI Processing Model and provide a tutorial on how to use TEI Publisher app to experiment and try out various ODDs containing processing model instructions, upload users’ own files and create a custom ODD, and, finally, generate their own, standalone edition using the App Generator. As an inspiration it will also present examples of real apps built with App Generator and other systems employing TEI Processing Model.

It is hoped that exposure to the concepts and technologies presented during the workshop will give its participants a point of exit in the task of publishing their own research data.

[1] eXist Solutions, Germany

Introduction to TEI in Japanese 日本語による TEI 入門

Kiyonori Nagasaki[1]
(Room A4)

This workshop aims to share concept usage of TEI for Japanese speakers. TEI guidelines have been the de facto standard for humanities research and data sharing in Western countries especially for linguistics, philology, and biography, providing various ways of description of textual structures. However, the guidelines have not yet known well so far. This workshop provides an opportunity to have basic knowledge of the TEI to Japanese participants who are interested in the TEI but not familiar with it.

Agenda:
Morning:

- Concepts of TEI with some use cases
- Introduction to TEI with Oxygen XML Editor

Afternoon:
How to treat a corpus encoded by TEI

- Encoding a modern novel “Hashire Meros”
- Encoding a scholarly edition of a Japanese classic “Tales of Genji”
- General discussion

Requirements:
A laptop computer which can connect WIFI and USB memory

If you are interested in encoding correspondence but not have enough knowledge of the TEI, it is recommended to participating this workshop in the morning to learn the basis of the TEI before the correspondence workshop.

本ワークショップでは、日本語話者に対して TEI の思想と初歩的な使い方を解説する。TEI は欧米の人文学資料を扱うにあたってのスタンダードとして、様々なデジタルアーカイブの裏側で利用され、様々な便利で高度なサービスを低コストで提供することを可能としてきている。特に言語学や文献学、書誌学においては様々に活用されてきており、碑文や手紙の情報を構造的に記録・記述することに関しても広く用いられるようになっている。それにも関わらず、日本ではまだ知名度は高くない。TEI が掲げる思想と実践についての知識が広まることは、日本の人文学とデジタルアーカイブとの関係をよりよいものにし、両者をともに深めていくことに大いに資すると期待されることから、日本語による TEI 入門ワークショップをここに提供する。内容の概略は以下の通りである。

午前

- TEI の思想：いくつかの事例紹介も含めつつ
- TEI の初歩：Oxygen XML Editor と TEI の基本的な構造

午後

- TEI で構造化された言語コーパスの分析に向けて
- 『走れメロス』の TEI による構造化
- 『源氏物語』池田亀鑑本の校訂情報の構造化
- まとめの議論

[1] International Institute for Digital Humanities, Japan

参加にあたって必要なもの

USB メモリを接続可能で Wifi 接続も可能なノートパソコン

留意事項

なお、特に手紙の構造的記述（手紙のメタデータ）に関心があるが TEI の初歩をまだ押さえてない、という場合には、午前中のみこのワークショップに参加し、午後は手紙のワークショップに参加するというやり方もあるのでご検討されたい。

[Day 1 (Sun) Morning] Half-day Workshops

Publishing TEI on the Web without XSLT: CETEIcean (/sɪˈti:ʃn/)

Raffaele Viglianti[1], Hugh Cayless[2]

(Room A1)

This half-day workshop will introduce CETEIcean (pronounced /sɪˈti:ʃn/, similarly to the word “cetacean”), a JavaScript library for displaying TEI in a web browser. Instead of sticking with the semantically poor element set of HTML, CETEIcean reframes an isomorphic transformation of TEI as HTML, by registering modified TEI elements with the browser using the new Custom Elements technology. Loading a TEI document with CETEIcean will make it fully usable by the browser: TEI elements can be styled with CSS and manipulated for interaction with JavaScript, just like HTML and thus avoiding XSLT transformation steps. CETEIcean’s approach is inspired by TEI Boilerplate (http://teiboilerplate.org/), but it is based on web standards and is more flexible. It does not require an in-browser XSLT step, nor any modification to the source XML; TEI content can be loaded in the browser via an AJAX call, or via server-side processing.

Motivations

Many scholarly publications powered by the TEI rely on server side infrastructure for publication, typically by leveraging XML technologies such as XSLT and XQuery (via XML databases). The capability and flexibility of these tools is undeniable, so why use a different approach?

- Web technologies for web publication. XSLT/XQuery are not sufficient, by themselves, to create a digital publication on the web, they are simply a means to transforming TEI into HTML. But to create a fully featured publication on the web, HTML needs to be supported by CSS for styling and JavaScript for user interaction. What if we could skip the transformation part and focus on the web technologies needed for a digital publication?
- Getting started faster. By focusing on web technologies, new TEI users will not have to learn XSLT as well as the TEI to publish their first document.
- Semantics. Converting TEI to HTML is the most common and most practical way of publishing TEI texts on the web, but HTML lacks what TEI has: a very well-considered and mature set of semantic tags for encoding texts. When converting TEI to HTML the semantic distinctions in the markup are often lost in favor of typographic distinctions in the display. In other words, the data model represented in the TEI fails to carry over to the online version.

[1] University of Maryland, USA
[2] Duke University, USA

- Preservation. XSLT/XQuery transformations are often performed "on the fly" by server-side technology to provide data to a front-end application (written in HTML, CSS, and JavaScript). By using CETEIcean, TEI can be built into "static sites" that do not require server-side computation, making it easier to preserve the application into the future.

Syllabus
This half-day workshop will cover the following topics.

- Introduction to CETEIcean and motivations.
- Using CETEIcean to build a static site from a provided HTML template. We will discuss the structure of the template and how to change it for future use.
- Using CSS to style CETEIcean TEI. No previous knowledge of CSS required.
- Adding user interactivity to your TEI via CSS and simple JavaScript. We will discuss a number of examples and address questions/requests from the attendees.
- Publishing on-line. We will look at simple ways of publishing a CETEIcean-powered site on-line for free via GitHub, DropBox, others.

Introduction to XPath

Syd Bauman[1], Sarah Stanley[2], Elisa Beshero-Bondar[3]
(Room A3)

Abstract
This workshop will cover the basics of XPath, an extremely powerful language for navigating and processing XML. With XPath, you can search for highly specific textual features, such as find all the elements with a specific @type value, or how often a certain character in a play speaks in verse or in prose. XPath is the underlying language used by XSLT, XQuery, Schematron, the TEI Processing Model, and in some XPointer schemes and TEI ODD constructs. While experience with writing valid TEI is desirable, participants only need to know how to write well-formed XML to benefit from this workshop.

Outline
This workshop will cover the basics of XPath, the syntax used for navigating and processing XML documents. Learning XPath will help any encoder of TEI to comprehend how their code fits and nests in the XML tree, and how to express relationships among its various parts. You can apply XPath to check for accuracy of text encoding and to identify patterns and irregularities in your coding. Learning XPath will also improve your vocabulary for describing the various units ("nodes") of XML that are navigable, and you will learn how these are "walkable" and tractable for processing. Writing XPath is a fundamental skill for the transformation and querying of XML documents, and for developing schemas to check your encoding. XPath is also used by the TEI community in various contexts, for adapting the TEI Processing Model and for designing TEI ODD customizations and TEI XPointer schemes.

We will not cover all the applications of XPath in this workshop, but participants will gain perspective and experience to continue practicing and learning. In this workshop, you will learn about the vocabulary that XPath uses to talk about the component parts of XML trees. You will also learn about XPath syntax, including how to find ancestors and descendants and how to refine queries with predicates. While familiarity with TEI is desirable, it is not absolutely necessary for this workshop. At minimum, however, we

[1] Northeastern University, USA
[2] Florida State University, USA
[3] University of Pittsburgh, USA

recommend that participants have written an XML document and have basic familiarity with the XML syntax and the concepts of well-formedness and validity before attending this workshop. If you have written code with angle brackets and want to understand how you can process it and what you can build with it, this workshop is for you.

During the course of the workshop, you will learn how XPath can help you transform data from your documents into various new formats and visualizations, and you will see some examples of how XPath is used in practice. The workshop will include a short demonstration of XSLT XQuery, and participants will get to look at how XPath is used as a basis for transforming documents. While the workshop will not cover how to write XSLT or XQuery, which are used to transform and process XML documents, you should leave with a good grasp of how XPath is utilized in these contexts and the workshop will prepare you for more advanced work in writing these kinds of programs.

This half-day workshop will consist of:

1. Introduction, setup, and basics of XPath vocabulary (**~1 hour**)
    a. Set up: (check oXygen installations, load files)
    b. Understanding XPath expressions:
        i. Navigation: XPath Axes
        ii. Path Steps
        iii. XML Node types
2. Introduction to XPath syntax (**~1 hour**)
    a. Constructing XPath queries
    b. Refining your searches with predicate “filters”
    c. XPath functions (a selection)
3. Hands on practice writing XPath expressions (**~1 hour**)
4. Future directions (**~1 hour**)
    a. Overview of some uses of XPath in the TEI
    b. Testing TEI with XPath (demonstration of Schematron)
    c. Transforming TEI with XPath (demonstration of XSLT)

Participation and requirements

To participate in this workshop, attendees must bring their own laptops. Participants will receive e-mail in advance of the workshop with information about installing the <oXygen/> XML Editor and be provided an extended complimentary trial license key courtesy of SyncroSoft. No other installation is necessary.

[Day 1 (Sun) Afternoon] Half-day Workshops

Introduction to TEI Encoding of Correspondence Meta Data

Anne Baillot[1], Stefan Dumont[2], Sabine Seifert[3], Peter Stadler[4]

(Room A3)

Workshop description

The objective of this training workshop is to convey the encoding of correspondence with the still rather new ‘correspDesc’ element. It will also present the derived Correspondence

[1] Le Mans Université, France
[2] Berlin-Brandenburg Academy of Sciences and Humanities, Germany
[3] University of Potsdam, Germany
[4] Paderborn University, Germany

Meta Data Interchange Format (CMIF) for interconnecting letter collections. The aim of this workshop is twofold: 1) the dissemination of the encoding possibilities and 2) to get feedback that may help to improve the current Guidelines (with regard to correspDesc) and which may result in best practice models of letter encoding.

Details

With the release 2.8.0 of the TEI Guidelines in 2015, several new elements were introduced especially for the encoding of correspondence (letters, postcards, emails, etc.) (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD44CD). The workshop will discuss those elements and how they can be applied to the participants' own correspondence collections. The focus will thus be on the meta data of correspondence material, separating the communicative, the material and the textual aspects which are to be encoded within 'correspDesc', 'sourceDesc' and 'profileDesc', respectively.

Additionally, the Correspondence Meta Data Interchange Format (CMIF) will be presented (https://github.com/TEI-Correspondence-SIG/CMIF). The CMIF is a constrained TEI customization provided by the Correspondence SIG for facilitating interchange of correspondence meta data which builds on authority controlled IDs (e.g. VIAF and GeoNames) for identifying entities, and the W3C format for dates. This enables e. g. the web service "correspSearch" (http://correspsearch.net/) to aggregate index listings of various letter collections and to enable searching across those collections. This web service also allows to record correspondence meta data in a CMIF file via the "CMIF creator" (https://correspsearch.net/creator/).

Participants are welcome to bring correspondence materials they might be aware of or working on and that can serve as a basis for discussion and hands-on encoding phases during the workshop. With these examples, we want to get feedback that may help improve the current TEI Guidelines with regard to the encoding of correspondence and that may result in more enhanced encoding examples (https://github.com/TEI-Correspondence-SIG/correspDesc) and best practice models.

References

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

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

Spoken Language: Tools and Workflow for Creating and Editing Data and Metadata

Carole Etienne[1], Christophe Parisse[2], Loïc Liégeois[3]
(Room B2,3)

Most part of people working on oral corpora have to deal with both the choice of a set of metadata clear enough to make their corpora reusable by a large community and the use of several transcription tools which require to develop specific software to make them working together with no lack of information.

[1] ICAR, France
[2] Modyco/INSERM, France
[3] LLF/CILLAC-ARPss, France

Generally, scripts have been developed in the different teams matching only some needs, working either on Windows, Mac, or Linux but rarely on more than one system, delivering data in a dedicated format difficult to share, not really user-friendly, with no real maintenance and evolution. In this workshop, we would like to present the two free and open-source solutions we have developed inside IRCOM/CORLI and ORTLOLANG two French infrastructures which are available for a large community:

- teiMeta is a tool for editing metadata in any XML (and TEI) file. Its goal is to allow editing or adding a common set of metadata to any file without damaging the other data in the file. The software is based on an ODD file and a stylesheet, and TEI is generated automatically. So it is possible to create as many versions of metadata editing as required for real applications. teiMeta works inside a web browser, so it is multi-system compatible.
- teiConvert is a tool for converting transcript file between different software (Transcriber, Clan, Praat or Elan) using a TEI pivot format. It can generate data in TEI format but also csv, txt (Utf 8), docx, txt for TXM software or txt for Trameur/Lexico software. teiConvert is written in Java, so it works on many systems. An alternate web interface exists for people that do not want to install Java and to use command line instructions.

After a presentation of teiMeta and teiConvert, we plan to organize a practical session where we can work on some different examples of oral corpora of different types: acquisition, sociolinguistic, phonetic, rare languages or interactions. The goal of the workshop is to show how to obtain high quality data in TEI to be used for research purposes, data sharing and preservation.

To make this practical time more efficient, people could send us in advance some examples of their metadata and transcripts to work on their own data and find solutions together. We can give some examples of different stylesheets or ODD files to adapt the tools to participant's needs.

References

teiConvert: http://ct3.ortolang.fr/teiconvert
teiMeta: http://ct3.ortolang.fr/teimeta

Liégeois, L., Etienne, C., Parisse, C., Benzitoun, C., Chanard, C. (in press), "Using the TEI as pivot format for oral and multimodal language corpora". *Journal of the Text Encoding Initiative*, 10, halshs-01357343.

Liégeois, L., Etienne, C., Benzitoun, C., Parisse, C. (2017). *Vers un format pivot commun pour la mutualisation, l'échange et l'analyse des corpus oraux*, Floral, Orléans.

Parisse, C., Benzitoun, C., Etienne, C., Liégeois, L. (2017) *Agrégation automatisée de corpus de français parlé. Journées de Linguistique de Corpus*, Grenoble, France.

[Day 2 (Mon) 16:45-19:00] Posters & Demonstration (RoomA 2,3,4)

TEI Technical Infrastructure

Hugh Cayless[1], James Cummings[2], Martin Holmes[3], Peter Stadler[4], Magdalena Turska[5]

The main goal of the Text Encoding Initiative (TEI) is the development of a standard for encoding of textual phenomena and manifests itself in the TEI Guidelines and the derived (formal) schemata. For the production and maintenance of these Guidelines and its various file formats, as well as for processing any kind of TEI document, a larger ecosystem of tools, methods, and technical service infrastructure have evolved around the TEI standard.

Since the most common serialization format for TEI documents is XML, a lot of generic XML tools are employed, e.g., Apache ANT as a build tool, Saxon as XML processor, and XSLT and XQuery as scripting and query languages. In fact, the most prominent piece of software that the TEI Consortium produces are the TEI Stylesheets, a set of XSL stylesheets that convert to and from TEI files. Supported import and export formats include docx, markdown, HTML, PDF, and many more. But the TEI Stylesheets not only convert 'regular' TEI documents but also TEI ODD customization files, acting as an ODD processor to produce both documentation and schemata from an ODD customization—again in various possible output formats.

The TEI stylesheets are at the core of most available TEI transformation services including the OxGarage RESTful web service, the oXygen TEI frameworks, the MEI Customization Service, and the TEI Jenkins continuous integration servers which test the TEI build process. Other services aim at providing an online editor for TEI ODD customizations rely on the aforementioned OxGarage services to handle these transformations e.g., produce JSON serializations of the TEI specifications for their internal data structures, and return schemata and documentation to the user.

The whole multitude of tools and services can best be understood by looking at the TEI Git repositories that are hosted at GitHub under https://github.com/TEIC. The most prominent are the TEI Guidelines and Stylesheets themselves, while CETEIcean is another rising star. CETEIcean is a pure CSS and Javascript renderer for TEI files which facilitates not only the display, but also the online editing of TEI documents. It is a smart front end library which can be used in one's own project by simply including the needed javascript and CSS files.

Most of the aforementioned tools depend on other services and software. It is the aim of the poster to illustrate these dependencies and to give a thorough overview of the tools and services (actively) maintained by the TEI Consortium. Such an overview would help to distinguish the various tools and services for those members of the TEI community, who find it hard to know what the relationships are between Roma, OxGarage, the Stylesheets, and the Guidelines. While the more tech savvy members would hopefully appreciate these insights for installing and running these services on their own hardware. Finally, the TEI Consortium itself would benefit from better documentation of their service architecture and the feedback of their users gained by this poster.

[1] Duke University, USA
[2] Newcastle University, UK
[3] University of Victoria, Canada
[4] Paderborn University, Germany
[5] eXist Solutions, Germany

@SameAs TEI? The Pitfalls and Perils of Creating Linked Data from TEI-Encoded Resources

Constance Crompton[1], Michelle Schwartz[2]

The renewed interest in the relationship between TEI and linked data (including the theme of this year's JADH conference, and the revitalization of both ADHO's LOD SIG and the TEI's Ontologies SIG) is a promising sign of our field's engagement with linked data, and our readiness to join international efforts to produce and publish that data (Ciotti and Tomasi; Huber et al.; Pattuelli et al.; Lehmann et al.; Shadbolt et al.; Hellmann et al). Currently, however, linked data only makes up 1% of the web (Simpson and Brown), very little of it coming from meticulously created and often peer-reviewed TEI-encoded data.

We propose a poster presentation introducing our decision tree for mapping TEI elements onto a number of existing ontologies. Our method for determining the meaning of any particular TEI element is twofold: first by reference to its definition and examples in the TEI Guidelines, and second, from its local use in one of our sample datasets, drawn from a number of long-standing TEI-encoded projects. Our current sample TEI-based data sets represent 45,000 entities including manuscripts, books, periodicals, biographies, art works, legislation, places, and events. The data spans four hundred years, two regions, five religions, and three languages, all with particular historical-contextual specificity. While our goal is to map all elements onto the SKOS, Schema.org, DBPedia ontology, DC, FRBRoo, BIBO, CIDOC-CRM, FOAF, CWRC, Open Annotation, GeoNames, or W3C Provenance ontologies, the process is far from a straightforward one-to-one mapping. We are especially interested in the poster format, as we are keen to solicit feedback from peers on the balance between granularity and generality in the representation of TEI-based information as linked data. We would be glad of the opportunity not only to meet with Ontologies and Tools SIGs, but also to ask other conference goers, where appropriate and if they are interested, to share snippets of their TEI, in order to broaden our understanding of the diversity of local TEI use.

References

Ciotti, Fabio, and Francesca Tomasi (2016) "Formal Ontologies, Linked Data, and TEI Semantics." *Journal of the Text Encoding Initiative*, no. Issue 9, *jtei.revues.org*, doi:10.4000/jtei.1480.

Hellmann, Sebastian, et al. (2014) "Knowledge Base Creation, Enrichment and Repair." *Linked Open Data-Creating Knowledge Out of Interlinked Data*, edited by Sören Auer et al., Springer International Publishing, pp. 45–69.

Huber, Jakob, et al.(2014) "LODE: Linking Digital Humanities Content to the Web of Data." *IEEE/ACM Joint Conference on Digital Librarie*s, http://arxiv.org/abs/1406.0216.

Lehmann, Jens, et al. (2012) "DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia." Semantic Web Journal, vol. 1, no. 5, pp. 1–29.

Pattuelli, M.Cristina, et al. (2013) "Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project." *The Code4Lib Journal*, no. 21, Code4Lib Journal, http://journal.code4lib.org/articles/8670.

Shadbolt, N., et al. (2012) "Linked Open Government Data: Lessons from Data.gov.uk." *IEEE Intelligent Systems*, vol. 27, no. 3, pp. 16–24. *IEEE Xplore*, doi:10.1109/MIS.2012.23.

Simpson, John, and Susan Brown (2014) "Inference and Linking of the Humanist's Semantic Web." *Implementing New Knowledge Environments*.

[1] University of Ottawa, Canada

[2] Ryerson University, Canada

What to Do with All the Witnesses? Challenges of a Multi-witness Iterative Digital Edition of John Denham's *Coopers Hill*

James Cummings[1], Tiago Sousa Garci[1]

The digital edition of John Denham's *Coopers Hil*l is trying a different approach to digital scholarly editing. It is collating the vast majority of known copies of the seventeenth-century poem (132 out of 149), and has created an infrastructure for the addition of the remaining witnesses in the future. Given the availability of digital surrogates, an initial version is being produced in a short period, with further witnesses being progressively added over time, dynamically altering the composition of the edition.

From an encoding point of view, the challenges behind this approach are twofold: how to manage -- that is, sensibly and rigorously encode while maintaining readability and modulary -- the large number of witnesses, and how to allow for the progressive addition of new witnesses without reconfiguring the entirety of the edition's TEI data model?

This poster will set out the challenges and solutions proposed for these questions, as well as the rationale behind them. The first challenge has implications for both editor and reader: the additional data has the potential to decisively transform the editorial work, but its volume makes traditional editorial methods impractical, making it a perfect candidate for computer-aided editorial work; for a reader, the challenge is in how to present the information intelligibly and concisely. The second challenge hinges on the first: the digital edition needs to allow for the progressive addition of new witnesses that might influence the editorial work without undermining the organisation of the edition itself. The challenges of *Coopers Hill* make it a perfect candidate for a stand-off markup approach, a much underutilised approach in digital scholarly editing, and this poster will serve as a case study for its advantages as a model for a multi-witness, iterative digital edition.

Keywords
digital editing, standoff markup, multi-witness, iterative development

A Markdown Approach for the Diffusion of TEI Usage in Classical Japanese Studies

Yuta Hashimoto[2]

Classical Japanese texts have various kinds of special notations that cannot be expressed in plain text such as *ruby, warigaki, okuriten, kaeriten*, and so on. However, it is still rare for scholars of Japanese studies to use TEI encoding to digitize classical Japanese texts, partly because there are only few scholars, librarians, and archivists familiar both with classical Japanese and XML syntax, and partly because paper-based academic publishing is still dominant in Japanese studies.

The ideal approach to promote the usage of TEI in Japanese studies is to organize large-scale training programs to teach XML syntax and TEI encoding, which will however take much time and labor. Another promising approach is to develop a mini language with a simple grammar that is easy to learn for those without knowledge on XML and can be automatically converted into XML and TEI. One example of such a language is

[1] Newcastle University, UK
[2] National Museum of Japanese History, Japan

Markdown[1], a light weight markup language with plain text formatting syntax, which was originally designed as an easy-to-write format to generate HTML.

Based on this idea, the author designed a mini language to encode classical Japanese texts in a simple way and has been developing a parser for the language which scans a source text, performs tokenization and lexical analysis on it, builds an Abstract Syntax Tree (AST), and generates a XML snippet that can be embedded in TEI documents. In this poster, the author will demonstrate his mini language and its parser program, showing how they makes it easy to encode classical Japanese texts.

Attempts to Describe Japanese Manuscripts

Hiroyuki Ikuura[2]

1. Introduction

In describing bibliographic information of classical texts, it is expected to use a bibliographic format that has high flexibility to allow documenting of various characteristics of the material as fully as possible. On the other hand, the bibliographic systems available today are being expected to interoperate as extensively as possible. Text Encoding Initiative guidelines are well-known to address both of these needs. This study discusses the way in which TEI guidelines can be adapted to the bibliographic information of Japanese classical texts, using the Toganoo Collection held by UCLA Library as a case study.

2. Materials used for the study

Consisting of classical and modern texts, the Toganoo Collection had held by Taganoo Shōun (1881-1953), a great scholar of Buddhism who served as president and chief librarian of Kōyasan University. The collection contains total 342 titles in 968 volumes and 2 scrolls. The purchase of this collection was initiated by Ashikaga Enshō, the 2nd chair of the Oriental Department of UCLA, in 1962-1963. In support of a growing interest in area studies in the US universities in the late half of 20th century when Japanese collections were among others that began to build in the libraries. UCLA Oriental Department (now Asian Languages and Cultures Department) needed to build a collection of Buddhist books to support its newly established course in Buddhism. The collection is one among those Japanese collections outside Japan that wait to be uncovered with detailed bibliographic information. It contains texts relevant to the historical studies of temples in Japan.

It is a collection of Buddhist texts that belong to a different genre among other pre-modern texts of Japan like Japanese literature. As such that this collection is selected as an appropriate resource to examine the applicability of TEI and identity specific issues in markup. Toganoo collection were selected during their research on Japanese pre-modern materials held by UCLA Library in the 1990s. However the bibliography omits detailed information including that of ownership such as signatures on the cover and collectors' seals that indicate the ownership history before they came into Shōun's possession.

3. Two preceding studies

We have a preceding case study which analyzes the application of TEI method of bibliographic description of Japanese classical texts. Kazushi Ohya "Unit-based Scheme Connection Between TEI and Original Scheme To Promote Data Sharing Beyond Cultual Diversities" (TEI conference 2014) points out TEI elements correspond inadequately to Japanese classical texts. The author concludes the structural difference between the two systems does not allow the adaption of TEI but suggests the use of TEI scheme as a reference to build a descriptive structure based on the current card data. It is reasonable

[1] John Gruber, Markdown, https://daringfireball.net/projects/markdown/.

[2] Waseda University, Japan

to suggest an automatic conversion of data from the bibliographic cards into the TEI structure as much as possible without altering the present structure of the cards because much of the card data has been already digitized.

The University of Cambridge Library's digital library uses TEI for the bibliographic description of classical text images including those of East Asian and allows online searching as well as downloading of the TEI files. This includes a record described with TEI markup by their Japanese librarian and a record of digital images created by Ritsumeikan University through their recent project. The detailed and structured bibliographic information including the history of ownership of these records allows searching of specific characteristics of an item. Downloaded TEI files can be used for machine processing, too. It has not been however verified whether these records addresses the structural differences pointed out in the study aforementioned.

4. Issues in the Markup of Japanese Classical Texts

The common problem in describing bibliographic information of Japanese classical texts in TEI is there is only one element that may correspond to multiple conventional bibliographic entries. This is above all immediately pointed out in the previous study, which attempted to match TEI elements against data entries of a bibliographic data card. Titles in Japanese classical texts appear in various places and combinations. As such that the element <title> alone is insufficient to describe them. Some lack naidai(caption title), others have the cover replaced after the completion of the text. Some works have multiple titles. In case of hanpon(published prints, mainly woodblock printing), they may have their titles in mikaeshi (verso of the cover), tobira(page after the cover page or title page), jo(preface), mokuroku(table of contents), hanshin(folding edge of the pouched leaves), etc. Adding attributes to the <title> element allows to describe multiple title entries. A specific attribute of a title may be described within "" of <title type= "">. In this study, several attributes are created such as <title type= "cover"> for gedai(a title written on the front cover or on a piece of paper), <title type= "caption"> for naidai, <title type= "alt"> for alternative titles.

A cover contains various critical information for Japanese classical texts such as the location of the title, color, design, and material of the cover. We can use attributes to describe these as in <bindingDesc><binding><decoNote type= "">.

Certain elements however allow no attributes. For example, the element <layoutDesc> contains no more than <layout> nor does it allow attributes by type= "" to describe layout information such as keikaisen(ruled lines on pages), jikō(the length of the first line of the text), and kyōkaku(frame of print, a frame that surrounds the body of text in each page).

A TEI markup file is considered to describe one classical text. This poses a problem in cases where multiple works are copied (gassha) in a text/book or multiple texts are bound into one text/book form (gattetsu).

5. Conclusion

Images open to public online without even basic textual bibliographic information require caution as research resource. As digital images become more available in large amounts, textual bibliographic information becomes more critical to support them. We need to ensure a flexible and interoperable bibliographic description system to accommodate this expansion of digital images. We have thus discussed enough TEI elements applicable to Japanese classical texts. There may still remain other elements useful to describe them. As such that we are aware of a need to further examine to expand applicable TEI elements.

quoteSalute: Integrate <salute> in Your Own Correspondence

Oliver Pohl[1], Stefan Dumont[1], Lou Klappenbach[1], Marvin Kullick[1], Frederike Neuber[1], Luisa Philipp[1]

quoteSalute[2,3] strives to make data of digital scholarly editions of letters (DSELs) accessible in a playful fashion by enabling users to integrate salutations from DSELs in their own email correspondence. At the moment quoteSalute aggregates around 750 salutations from various sources including salutations by the German scholar Wilhelm von Humboldt, the romantic writer Jean Paul, the female writer and countess Erdmuthe Benigna von Reuß-Ebersdorf and many more.

Nowadays people often close their letters and emails using uninspired phrases like “Kind regards” or “Best”. However, data in DSELs show that this has not always been the case. In the past, adding a personal salutation to your own correspondence was very common. From a scholarly stance, salutations provide important insight about the relationship between the writers of a correspondence.

The foundation of quoteSalute is a curated TEI-XML text corpus which has been created by extracting <salute>-tags[4] from TEI-XML-encoded DSELs. For providing users with fitting salutations, we annotated the data adding the language of each entry, the intended gender of both sender and receiver, and whether the salutations should be used in formal situations or rather colloquially.

When visiting the project page, users are presented with a random salutation. They can then copy the salutations and paste it into their own emails. Additionally, a Thunderbird extension[5] is already in development.

quoteSalute demonstrates how the editorial work on DSELs can be integrated into everyday emailing. Furthermore, if the application is used more widely, the editions themselves are likely to enjoy greater recognition beyond the boundaries of the humanities, as each quotation links to the letter of the respective edition. In the context of the TEI-community, we hope to encourage other projects to use the <salute>-tag more consequently as well as to add their own salutations to the quoteSalute-corpus[6][7][8].

---

[1] Berlin-Brandenburg Academy of Sciences and Humanities, Germany

[2] quoteSalute: Inspiring greetings for your correspondence. 19th April 2018. https://correspsearch.net/quotesalute/index.xql

[3] GitHub-Repository quoteSalute. 19th April 2018. https://github.com/telota/quoteSalute

[4] TEI element salute (salutation). Version 3.3.0, 31st January 2018. http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-salute.html

[5] GitHub-Repository quoteSalute_xpi (Thunderbird extension). 19th April 2018. https://github.com/telota/quotesalute_xpi

[6] quoteSalute documentation (currently only in German). 19th April 2018. https://correspsearch.net/quotesalute/index.xql?id=doc&l=de

[7] quoteSalute contributor templates. 19th April 2018. https://github.com/telota/quoteSalute/blob/master/doc/template.xml

[8] quoteSalte example transformations for contributors. 19th April 2018. https://github.com/telota/quoteSalute/tree/master/doc/xslt-examples

The Standardization Survival Kit: Make your Arts and Humanities Research Go Standard - TEI Inside!

Marie Pure[1], Charles Riondet[1], Laurent Romary[1], Dorian Seillier[1], Lionel Tadjou[1]

The Arts and Humanities research community has to address new challenges raised by the increasing amount of digital sources, contents and tools. Digital infrastructure project, such as PARTHENOS [2] (in close collaboration with the CLARIN and DARIAH infrastructures), aim at supporting this turn by offering innovative solutions to connect digital tools and contents to Arts and Humanities researchers' needs. One of these solutions is the Standardization Survival Kit ("SSK"), a platform to help humanities scholars understand the crucial role that proper data modelling and standards have to play in making digital contents sustainable, interoperable and reusable.

The SSK is an overlay platform promoting a wider use of standards within Arts and Humanities. It aims at providing documentation and resources concerning standards. The interdisciplinary nature of the TEI obviously makes it a central reference source for several research scenarios described in the platform.

Moreover, we went one step further to be consistent with our principles down to the details. We used the TEI as the actual SSK underlying data model for describing all parameterisable content, adopting and customizing the <event> element. The Standardization survival kit data is a collection of TEI files representing research processes, that we call scenarios, divided in steps. We considered that such a process, a mixture of intellectual and concrete actions forming a methodological sequence, could be represented as a series of events. A scenario is a list of events (<listEvent>) and the steps are events (<event>). To maximize the reuse and the customization of the scenarios, the scenarios and the steps are represented in separate files (see here a scenario file and here a step file, available in GitHub), connected by pointers. Inside events, bibliographical resources are referenced in <ref> elements, put together in a <linkGrp>, pointing to a Zotero database.

The SSK beta version can be browsed here:
https://sskapplication.parthenos.d4science.org/ssk/#/

Semantic Minimal Retrospective Digitization of Edited Correspondence

Klaus Rettinghaus[3]

Correspondence is in general an essential key to research in many aspects. That is why around the world so much effort goes into editing corpora of letters. A problem is that most of the available edited correspondence is still imprisoned between cover boards and thus may stay hidden for the researchers area of interest.

But a full retrospective digitization is time-consuming and expensive, and for newer publications even prohibited by copyright laws.

The pure metadata for the edited correspondence however is not direct subject to copyright restrictions. And with the relatively new elements for *correspondence description,* which were introduced in the version 2.8.0 of the TEI Guidelines in April 2015, it is easily possible to encode this correspondence metadata in a structured and

[1] Inria, France
[2] https://cordis.europa.eu/project/rcn/194932_en.html
[3] Saxon Academy of Sciences in Leipzig, Germany

standardized TEI file, which then can be used to make this data interchangeable and thus easily searchable with other services like correspSearch.net.

The poster will discuss the mentioned possible problems in retrospective digitization in general, propose possible work flows and tools for the creation of a so-called *semantic minimal retrospective digitizatio*n, and the overall benefit of open sourced correspondence metadata.

The Digital Scholarly Edition of Andreas Okopenko's Diaries – Showcases and Research Questions

Laura Tezarek[1]

The estate of the Austrian neo-avant-garde writer Andreas Okopenko (1930–2010) is a comprehensive collection of diary notations, typescripts, notes, correspondence and other documents and was acquired by the Austrian National Library in 2012. Regarding text-genetical, poetological, (literary) historical and sociological aspects, the diaries are the core of Okopenko's estate.

Throughout his life, Okopenko, a pioneer in hyper-text literature, turned out to be a reserved but rigorous observer of the literary scene in post-war Austria. Up to now, little attention has been devoted to the impact he had on Austrian "experimental" literature during the early 1950s, where he assembled Vienna's most important progressive writers and can be considered a "node" in the network of the early Austrian neo-avant-garde. Therefore, the diaries in particular will provide a new perspective on the formation and the local value of Austrian Literature. Okopenko's perceptions and awareness of his own historical context make him a distinguished "archaeologist", "archivist" and "chronicler" of his time. As such, the estate is an impressive individualistic testimony of contemporary Austrian history.

The digital scholarly edition of Andreas Okopenko's diaries will be published in 2018 via the new edition platform of the Austrian National Library. A close collaboration between the Department of German Studies of the University of Vienna and the Literary Archives of the Austrian National Library guarantees an optimal valorisation of the estate and the adherence to standards in digital scholarly editing (using TEI).

This poster will present a selection of exemplary showcases of this digital scholarly edition, including spatial-temporal relations of data and source collections (DARIAH Geo-Browser) and their correlations as well as network analyses in regard to Austrian avant-garde literature circles.

"Coffee" and Other Coding Challenges: Lessons Learned through David Livingstone's Manuscripts in South Africa (1843-1872)

Mary Borgo Ton[2]

This poster describes the efforts of Livingstone Online's global team to develop a TEI-encoded critical edition of Dr. David Livingstone's manuscripts in South African repositories (http://livingstoneonline.org/in-his-own-words/livingstone-s-manuscripts-in-south-africa-1843-1872). For this new project, the team applied the TEI guidelines used in previous facets of Livingstone Online and developed over a 10-year period, including Livingstone's Final Manuscripts (1865-73) as well as multispectral critical editions of the 1870 and 1871 Field Diaries. The edition of Livingstone's Manuscripts in South Africa,

[1] University of Vienna, Austria
[2] Indiana University, USA

published in February 2018, reflects a shift in Livingstone Online's focus from its early emphasis on on Livingstone's place within western medical discourse to its efforts to foreground the African physical and social ecosystems that shaped David Livingstone's Victorian-era travels. Working with archival documents from African collections challenged us to rethink our coding practices, particularly the was that we use the element "term" and attribute "type," to take account of on-the-ground realities of African ethnic groups, organizations, occupations, plants, and foodstuffs. Livingstone's references to "coffee", for example, challenged the assumptions that we were making about the relationship between plants and the place where Livingstone writes about them, for he requested coffee grown in Mozambique to enjoy as he traveled throughout southern Africa. Our coding choices for terms like coffee reflect the difficulties of representing nineteenth-century African ecosystems, trade networks, and cultures, particularly since many individuals that Livingstone encountered in Africa where themselves not local residents but from other African regions, the Arabian Peninsula, or India. As Livingstone Online develops closer ties with African scholars, archivists, and institutions, we are eager to revise our coding practices to further foreground Livingstone's African contexts and to better represent the fluidity of nineteenth-century African cultures and material realities.

Towards a Rich Interface for ODD Customization: The New JavaScript Version of Roma

Raffaele Viglianti[1], TEI Technical Council

The TEI module for writing or customizing an XML markup language (the "tagdocs" module of chapter 22, "Documentation Elements"), and moreover a file defining or customizing an XML language using these elements, is typically referred to as One-Document-Does-it-all or ODD. Although ODD is a part of the TEI that is explicitly not about the encoding of humanities source material, it is an essential part of the lifecycle of many TEI projects and of the TEI itself. The procedural nature of these elements makes it possible to develop form-based user interfaces to generate ODD; indeed, the current tool for customizing the TEI schema, Roma (http://www.tei-c.org/Roma/), is widely used by the TEI community to generate project-specific schemata. Roma, however, does not fully reflect the possibilities of the ODD subset of elements, and has been afflicted by a number of bugs and issues, many of which remain unresolved and are too complex to be fixed by the TEI Technical Council.

The TEI Technical Council has opted to create a replacement for Roma from scratch, using modern web technologies, expanding the number of features supported, and focusing on user interface. This effort is led by council member Raffaele Viglianti, who is developing a web application (temporarily named "RomaJS") using the popular front-end libraries React and Redux. The full council provides regular feedback on both architectural decisions and the user interface. Just like the current version of Roma, this new version relies on the TEI Stylesheets to perform ODD transformations, including from ODD to XML schema formats. To perform these transformations, RomaJS interacts with OxGarage (http://www.tei-c.org/oxgarage/), an on-line service for TEI transformations.

This poster will introduce motivations for developing RomaJS, describe its architecture, and demo its features.

[1] University of Maryland, USA

Gaiji from Cradle to Grave: Encoding Traceable Identity of Characters in TEI

WANG Yifan[1]

Ideally, in TEI, which assumes a comprehensive character set, digitally unavailable characters are considered sporadic and/or exceptional. However, despite the active and continuous work on standardization of unencoded Han (Chinese) characters into Unicode and UCS, the queue of characters awaiting encoding will not likely to be cleared up in the foreseeable future. It means that reference to non-standardized characters (*gaiji*) is going to remain as an almost inevitable operation when conducting digitalization of an East Asian, especially pre-Modern, document using TEI. Such large-scale text database usually contains a proportional size of *gaiji* stock, most of which are expected to be migrated to the standard over time, gradually, but unevenly: “due to the way Unicode has been defined” (TEI P5 Guidelines), encoding proposals would have various outcomes—from assignment to a single code point, to rejection from inclusion—often in unintended form and indeterminate period for their proposers.

Since a considerable number of *gaiji* have a rather fluid station as standardization candidate, ensuring their diachronic traceability—making their history and current status in standardization process available—would be informative to human users and application software, contribute to long-term stability and usefulness of digital editions with *gaiji* module, and augment TEI feasibility towards East Asian documents that encourages TEI usage in this region.

In this presentation, we would like to discuss method and implementation which enable to provide character information related to standardization as metadata in the way capable to be incorporated into TEI framework, as an extension to *gaiji* module. It will encompass basic concepts regarding standardization process and possible solutions on element structure, description, synchronization, glyph identity, character relevance and other aspects.

[Demo] CSV2CMI – A Tool for Creating a Correspondence Metadata Interchange Format File

Klaus Rettinghaus[2]

With the relatively new elements for correspondence description, which were introduced in the version 2.8.0 of the TEI Guidelines in April 2015, it is now easily possible to encode correspondence metadata in a structured and standardized file. Hand coding whatsoever can be a bit of a hassle, especially with bigger corpora. CSV2CMI allows you to enter your metadata conveniently in a table, and transform it into a TEI file that complies with the CMI Format, which then can be used to make your data interchangeable and thus easily searchable with other services like correspSearch.net.
This means a great deal when it comes to letter corpora that have been published in the pre-TEI era or in making scholarly correspondence data more accessible.

The demonstration will present the scope of the tool, show examples where it has been used so far, and discuss future prospects.

The Tool is available on GitHub (https://github.com/saw-leipzig/csv2cmi).

[1] The University of Tokyo / International Institute for Digital Humanities, Japan
[2] Saxon Academy of Sciences in Leipzig, Germany

[Day 3 (Tue) Morning] SIG Sessions

9:00-12:15 (Half-day)

[Room A3] Ontologies (Constance Crompton)
[Room A4] East Asian / Japanese (A. Charles Muller, Kiyonori Nagasaki, Satoru Nakamura, Kazuhiro Okada)

9:00-10:30

[Room B1] Indic Texts (Patrick McAllister)
[Room B2,3] SIG Correspondence (Sabine Seifert)

10:45-12:15

[Room A2] Manuscripts SIG (Gerrit Brüning)
[Room B1] Text & Graphics SIG (Martin de al Iglesia)
[Boom B2,3] TEI for Linguists (Piotr Banski, Andreas Witt)

	9:00-10:30	10:45-12:15
A2	(JADH Session)	**Manuscripts SIG** (Gerrit Brüning)
A3	**Ontologies** (Constance Crompton)	
A4	**East Asian / Japanese** (A. Charles Muller, Kiyonori Nagasaki, Satoru Nakamura, Kazuhiro Okada)	
B1	**Indic Texts** (Patrick McAllister)	**Text & Graphics SIG** (Martin de al Iglesia)
B2,3	**SIG Correspondenc**e (Sabine Seifert)	**TEI for Linguists** (Piotr Banski, Andreas Witt)

[Day 3 (Tue) 13:45-15:15] LP Session I (Room A3)
Digital Archives

Encoding for Black Lives: Affordances and Challenges of TEI for Liberatory Archives

Jessica H. Lu[1], Raffaele Viglianti[1]

As more people live their lives online, demand for open access to historical documents and archives has increased. Text encoding provides one avenue toward digitization, especially as it allows us to embed interpretation into digital surrogates and build digital projects for public audiences. However, these opportunities also highlight questions about how TEI provides for the ever-shifting needs of diverse communities around the world.

In the U.S., African Americans are increasingly reliant upon digital spaces to preserve and share material that defies institutional archives; challenges problematic accounts of Black life; and pursues economic, social, and political liberation for Black people. Both The People's Archive of Police Violence in Cleveland and the Anti-Eviction Mapping Project serve as exemplars committed to creating and holding space for Black people whose lives have been systematically erased, silenced, or taken. This paper presents a corpus of contemporary American discourse about Black freedom as a means of discussing the affordances and challenges of TEI as a viable tool not for simply creating and maintaining digital archives, but for advancing projects that can uphold those same principles of justice and liberation. Particular TEI elements (even seemingly innocuous ones such as respStmt, profileDesc, pubPlace, persName, and personGrp) can be leveraged to embed and practice critical Black thought—theorizing by such preeminent scholars as Saidiya Hartman, Katherine McKittrick, Simone Browne, Achille Mbembe, and Christina Sharpe—within scholarly markup. Thus, we can use TEI not merely to encode and model original material, but to purposefully amplify perspectives of imperiled people, analyze texts in new ways, and exercise greater care in the representation of texts and the communities who produce them. Thinking beyond form and description, however, requires deliberate attention to the ways in which Black people are described as less than human or as non-persons; are silenced in dominant narratives; and are constantly (re)inventing and innovating language practices. These realities present exciting possibilities for those who utilize TEI markup to build liberatory archives, and those who engage with them.

Ultimately, this paper proposes a potentially productive relationship between TEI and Black digital work, and further suggests that TEI's role in the movement towards digital preservation hinges upon encoders' attention to human needs as well as scholarly aims. TEI's future as a global language relies upon its ability to evolve in conversation with the complex cultures and histories of communities around the world.

[1] University of Maryland, USA

REED:London and CWRC: The Digital Ecology of the Records of Early English Drama Project Integrating with Canadian Writing Research Collaboratory

James Cummings[1], Diane Jakacki[2], Susan Brown[3], Kimberley Martin[3], Carolyn Black[4]

The Records of Early English Drama (REED http://reed.utoronto.ca/) project over the last forty years has worked to locate, transcribe, and edit historical documents containing evidence of drama, secular music, and other communal entertainment and ceremony from the Middle Ages until 1642, when the Puritans closed the London theatres. Along with twenty-seven collections of records in print, the REED project is now publishing its new collections as freely available digital resources for research and education. At time of writing two new collections, marked up in TEI, have been published in digital form with several more on the way (see http://ereed.library.utoronto.ca/). The REED project has created a TEI ODD customization constraining the overall TEI scheme and enforcing project-specific rules as part of its new workflow. As new collections are released, existing digital materials are being integrated into this system. As part of a Mellon-funded project REED:London is undertaking TEI encoding of specific legacy print collections and integrating these and REED's named-entity references with the ontology and infrastructure provided by the Canadian Writing Research Collaboratory (CWRC see http://cwrc.ca/ and http://beta.cwrc.ca/reed). The REED:London project in its collaboration with CWRC is building a stable, extensible publication and editing environment that leverages the richness of the TEI source documents and integrates and expands the CWRC ontology, while supporting existing REED editors. This paper discusses how the TEI ODD customization fits into the REED workflow, the storage of named entity metadata, and the collaboration with CWRC as part of the REED:London project as an example of how a system of remediating information from manuscript into print can in turn be remediated through TEI into a new information ecology that unleashes the potential of interlinking within and beyond the REED project itself.

Bibliotheques françoises as a Virtual Workshop for the Literary History of Early Modern France

Toshinori Uetani[5], Guillaume Porte[6], Mathieu Duboc[5], Sandrine Breuil[5]

The Bibliothèques Virtuelles Humanistes (BVH: http://www.bvh.univ-tours.fr/) is a project run since 2002 by a research team founded by Marie-Luce Demonet in the Centre d'Études Supérieures de la Renaissance (CESR: the University of Tours and the CNRS, France) and now directed by Chiara Lastraioli. The BVH has been developing the *Bibliotheques françoises* project since 2015: a digital edition in XML-TEI and a data base of the first dictionaries of French authors, *Bibliotheques* of La Croix du Maine (Paris, 1584) and of Du Verdier (Lyon, 1585). Since the last TEI meeting (Vienna, September 2016), text and bio-bibliographical data base of the authors whose first name begins with letters A, B and C, of La Croix du Maine, are available in a beta version (http://bibfr.bvh.univ-tours.fr/). In this

[1] Newcastle University, UK
[2] Bucknell University, USA
[3] University of Guelph, Canada
[4] University of Toronto, Canada
[5] Centre d'études supérieures de la Renaissance (Université de Tours / CNRS, UMR 7323), France
[6] Université de Strasbourg, France

first phase, biographical data is linked to online references like viaf or data.bnf.fr., and bibliographical data to USTC (Universal Short Title Catalog, University of St Andrews) recordings, so that web users can even read digitized copies of some items online.

More than a hundred surviving copies of the *Bibliotheques* contain in them handwritten notes by their previous owners. One such edition recalls the memories of authors of the owner's acquaintance. Others note additional bibliographical information in the margin. Physical volumes of the *Bibliotheques* become thus a personal storage device of complementary data. Studied together, these additional notes constitute an important source of information on the Republic of Letters of Early Modern France. In a sense, the XVIIIth century collective edition of two *Bibliotheques* by Rigoley de Juvigny (Paris, 1772-1773) is an incarnation of these collaborative efforts for French literary history.

While the indexation of bio-bibliographical data of both *Bibliotheques* continues (books of author's initials A to H are now available), the second phase of this project aims to develop its encoding model for integration of layers of these manuscript annotations with their proper metadata, so as to show the collective efforts for information on French Renaissance authors, their networks and communities. For each entry of the *Bibliotheques*, a new system has to manage notes from dozens or hundreds of different copies and to enable modern researchers to comment them so as to create an open collaborative space of Literary History of Early Modern France.

[Day 3 (Tue) 13:45-15:15] SP Session I (Room A4)

The Italian Reception of the English Novel: A Digital Enquiry on Eighteenth Century Literary

Andrea Penso[1]

This short paper aims at showing the first results of the ongoing collaborative research project The reception of the English novel in the Italian literary press between 1700 and 1830: a transcultural enquiry into the early shaping of the modern Italian literary and cultural identity. The project, conducted between the Vrije Universiteit Brussel (VUB, Belgium) and the University of Guelph (Canada) investigates the reception of English novels in the Italian literary press during the Long Eighteenth Century (1700-1830). The analysis focuses on an existing corpus of data relative to the publication, dissemination, translations, critical reviews, and editorial advertisements of English novels in Italian literary newspapers and journals of the time. The main purpose of the project is to uncover how the English novels were introduced to the Italian readership through literary journalism with the application of Digital Humanities methodologies of investigation. One of the project goals is in fact to create a methodological paradigm that may be extended to the study of the reception of English novels in the literary journalism of other national traditions. The present paper therefore has three primary objects:

a) To show for the first time to the public the first research output: an open access, bilingual, and annotated digital repository, which consists of a Drupal-based software for corpora, and represents an immediate way to develop the research. The first step of the project has been the cataloguing, analysis, and digitization of the corpus of reviews. This preliminarily created digital database allows the subsequent computational, textual and critical surveys.

b) To illustrate the two main lines of approach that will be applied in order to digitally explore the corpus. The first consists of a stylistic and linguistic analysis of the reviews, which will be pursued equalizing and comparing stylistic and lexical constellations belonging to different discursive practices from a number of periodicals and journalist. Digital stylometry, word frequency and statistical analyses tools such as R, MiniTab and Intelligent Archive will be used during this phase. The study of the readers' response to the contents, spread by the novels via the reviews, is deeply connected to the stylistic analysis of the reviews. In fact, the outlining of the reviews' stylistic features is crucial to understanding in which ways the contents were revealed to the public, and how the audience was influenced in the perception of the moral values and the social messages of the novels. The second line of approach will concern the spatial analysis of the data, which will be mapped thanks to GIS (Geographical Information System) digital tools integrated with Geo-criticism. The analysis spatial analysis allows the visualization of popular reading trends in 18th and early 19th century Italy.

c) To show the applications of the TEI to the analysis of the corpus. The text encoding of the reviews, conducted following the TEI standards, makes possible – among other things – to identify the original and innovative elements of those Italian reviews that were based extensively upon foreign reviews. My paper will focus on this specific aspect: in fact, the encoding of the reviews allows to understand and visualise their "genealogical dimension" (i.e. the comparative analysis of reviews that were taken from French or English periodicals and made their way into the Italian press) and to explore the cultural specificity of the Italian journalistic practices. The 'genealogical' sources have already been identified, and the comparison between the encoded versions will allow to understand the extent of the influence French and English journalism had on the Italian press, and to outline the

[1] Vrije Universiteit Brussel, Belgium

specific Italian input. The TEI will therefore make possible to focus on the re-interpretations of Italian reviewers who drew on the Italian literary tradition but challenged its subjects, genres and linguistic structures. Ultimately, the TEI will also be applied to integrate the stylometric analysis and the gathering of Geospatial information: the short presentation will allow me to show a case study and to explain how the mark-up process proves to be a fundamental part of the methodology for the analysis of the corpus.

The Future of TEI and Events: The Case of a Register for Early Modern Sermons

Thomas W Dabbs[1]

This short paper is a continuation of a talk delivered at the DH2017 conference at McGill University in Montreal entitled 'The Extended Language of Religious Reform'. This paper will focus on the problem of encoding TEI/XML for events rather than full texts. The same work sample will be examined, a project to digitalize a register of sermons delivered during the 16th and 17th centuries in London.

These sermons are largely non-extant, but by their titles and event markers track the movement of English Reformation thought on the street, so to speak, and are an ignored historical cluster that, properly digitalized, would help scholars to envision precisely how radical religious ideas entered into a public space in London in the 16th century. These ideas, drawn from inauspicious beginnings, then progressed beyond London and England to power religious thought and political policy in the Americas and throughout the globe.

Aside from a brief overview of this project, this talk will focus on TEI/XML and the advantages or disadvantages that TEI might have as a platform beyond full text encoding and into metadata and ontologies. Three simple questions with complex answers will be considered: Is TEI currently a recommendable platform for a register of events? If not, is TEI being developed in such a way that it can be the preferred language for digital event platforms? Finally are such platforms as Schema.org, or Motools or LODE currently better than TEI for use in developing event ontologies?

Oral Literatures and Oral Musics: Their Analyzes and Interpretations Developed through the Use of TEI and MEI Encodings

Sylvaine Leblond Martin[2], **Henri Hudrisier**[3], **Mokhtar Ben Henda**[4]

The digital encoding standards TEI and MEI (Music Encoding Initiative) are representative of new forms of language through their ***descriptive functions and formalizations, exhaustives***, of the literary and musical texts, which consequently support a type of original research combining interpretation and subtle analysis.

Concerning ***oral, literary and musical documents***, this potential for adaptability of encoding is particularly useful. Indeed, writing and recording a literature or music of oral tradition presupposes an implicit, prior, underlying but ***unformalized analysis***. However, TEI and MEI oblige the user to describe the various stages of the progression of his decisions and thus to ***justify*** the choice of his interpretations. With regard to oral corpora, this type of ***explicit*** approach represents the advantage of constructing an analytical thought that the very notion of improvisation, inherent to oral literatures and musics, tends

[1] Aoyama Gakuin University, Japan
[2] Maison des Sciences de l'Homme Paris Nord and Chaire Unesco ITEN "Innovation, Transmission, Edition Numériques", France
[3] Membre de la Chaire Unesco ITEN, France
[4] Université Bordeaux 3 - Montaigne, Membre de la Chaire Unesco ITEN, France

to exclude: the musical improviser, and the literary storyteller, like magicians, never wish to reveal their secrets and spontaneously are fleeing most analytical attempts.

However, despite the main feature of literatures and oral musics which is to exist forcefully ***outside*** the traditional systems of writing, we now readily recognize the fundamental and efficient contribution of the writing and notation in the effort of ***conserving*** this "intangible" heritage. Moreover, the digital encodings of those literary and musical notations favor a work of comprehension of the texts thanks to the varied possibilities of forms of analysis and interpretations, which are part of the operation of the digital encodings.

This organization of digital standards not only reinforces the conservation action of literary and musical corpora of oral tradition, but also allows a thorough research of the mechanisms of ***conception*** and ***development*** at the base of these literary, musical, cultural and artistic creations. Thus, these ***intelligent markup languages*** TEI and MEI can express differences, such as those that exist between the joint actions of ***orality*** (transmission by voice) and ***aurality*** (listening by the ear) which establish two relations to the individual that exist simultaneously in the action of literary and musical oral transmission.

Nineteenth-Century Knowledge Project

Peter Melville Logan[1]

This talk will discuss the progress of a three-year-old project to build an extensive, open, digital collection for studying the structure of nineteenth-century knowledge by creating usable text from historic editions of the *Encyclopedia Britannica*. Today, we readily recognize a pervasive bias in the Eurocentrism of these entries, among other flaws. But at the time, the *Britannica* editions were the most authoritative comprehensive representation in the English-speaking world of knowledge as a whole. Knowledge has changed not just since that time but also during the publication of this material, from 1790-1911. This data set documents those changes. The goal of this project is identify patterns in the transformation of knowledge by mining the final data set. All of these works are available on the web, but their textual data is too inaccurate for valid text mining. This project thus creates the first accurate TEI edition of this valuable resource. The full corpus consists of 100,000 articles derived from 80,000 print pages. The TEI will be supplement with metadata using Named Entity Recognition. The Metadata Research Center at Drexel University will further enrich the data by adding subject metadata from both current and historical vocabularies, using an automated recognition program they developed. When complete, all individual entries will be made freely available through the Oxford Text Archive. It will be uploaded for other researchers in bulk form to the CORE Repository of the Humanities Commons.

Project URL: https://tu-plogan.github.io/

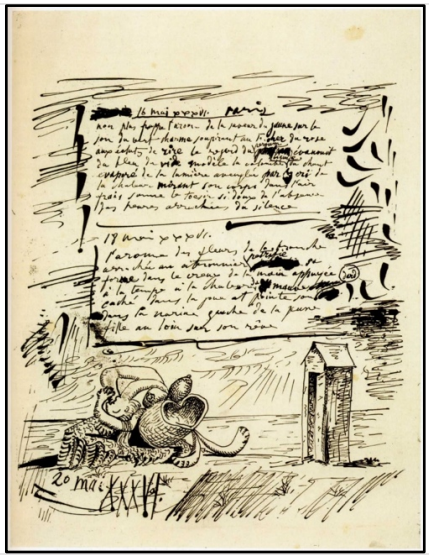
Encoding the Graphic Elements in Picasso's Poetry

Luis Meneses[2], **Enrique Mallen**[3]

At the TEI 2017 conference, we discussed how Picasso's complex poetry needed to be addressed in the pedagogy and practice of the Text Encoding Initiative (Meneses and Mallen 2017). In this abstract, we explore additional challenges pertaining to the encoding of graphic elements in Picasso's poems. The combination of text and graphic elements is

[1] Temple University, USA
[2] University of Victoria, Canada
[3] Sam Houston State University, USA

reminiscent of Picasso's collages in which different media were used on a single composition, as shown in figures 1 and 2.



Figures 1 and 2: Examples of the graphic elements in Picasso's poetry.
P. Picasso, "si yo fuera afuera ... (2)", Claude Ruiz-Picasso Collection, 1935.
P. Picasso, "non plus frappe ... (4), l'arôme des fleurs ... (4) (Baigneuse à la cabine)", Galerie Kornfeld und Cie, Bern., 1936.

Cox proposes that a broad definition of collage is "the incorporation of foreign elements into painting in a spirit of extreme juxtaposition ... so that the material character of representation leaps to our attention" (Cox 2010). To that extent, one could argue that collages ultimately serve as "seeing-machines", forming the testing-ground for the dialogue between material and sign. Along these lines, we propose that graphic elements in Picasso's texts also serve to direct readers' attention to the essential nature of words as "floating signifiers" (Derrida 2001), thus extending the semantics of the poems.

Picasso's poetry may be considered a Visually Complex Document, which as in the different "planes of consistency" of collages, present distinct layers of text and images that constitute integral parts of the document's representation (Audenaert 2008). To account for similar cases, Haaf and Thomas created a pure TEI subset for the unambiguous annotation of manuscripts (Haaf and Thomas 2016). However, we believe that Picasso's case is more complex, as it is hard to define the boundaries between graphics and text. The challenge that we face is how to encode elements in a relational manner in order to analyze possible ways in which graphics contribute to the reading experience. This duality is precisely what makes Picasso's poetry interesting from an encoding perspective. Our analysis proposes possible solutions for encoding these documents in TEI.

References

Audenaert, Neal (2008) "Patterns of Analysis: Supporting Exploratory Analysis and Understanding of Visually Complex Documents." IEEE Technical Committee on Digital Libraries. http://www.ieee-tcdl.org/Bulletin/v4n2/audenaert/audenaert.html.

Cox, Neil (2010) *The Picasso Book. First edition*. London: New York: Tate.

Derrida, Jacques (2001) *Writing and Difference*. 2nd Revised ed. edition. London: Routledge.

Haaf, Susanne, and Christian Thomas (2016) "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format." *Journal of the Text Encoding Initiative*, no. Issue 10 (December). https://doi.org/10.4000/jtei.1650.

Meneses, Luis, and Enrique Mallen (2017) "Visual Text: Encoding Challenges in Picasso's Poetry." presented at the TEI 2017, Victoria, BC, Canada, November 11.

[Day 3 (Tue) 15:30-17:00] LP Session II (Room A3)
New Teaching, New Methods

New Developments in the Guidelines: The att.linguistic Class

Piotr Bański[1], Susanne Haaf[2], Martin Mueller[3], Andreas Witt[4]

The topic of the perceived inadequacy of the vanilla TEI recommendations for building technologically effective, rather than merely theoretically "nice" linguistic resources has stirred some of the TEI community for years, essentially ever since the massively successful fork of the early TEI called CES (Corpus Encoding Standard), and later XCES and its variants have dominated the scene of language resources for a long time.

In January 2018, responding to an initiative by the Special Interest Group "TEI for Linguists" (LingSIG), the Guidelines introduced the mechanism of "lightweight grammatical annotation", in the form of the new attribute class, att.linguistic. The goals of the new class are twofold and addressing two groups of users. For language technologists, it opens a way to create effective linguistic resources that at the same time can use the full repertoire of the Guidelines for the purpose of identifying other features of text. For maintainers of existing non-linguistic TEI resources, philologists, historians and other DH scholars, the added value comes from enriching the existing structural markup with linguistic information, because this information might improve even discipline-specific tasks such as the identification of names of persons or places, or also authorship attribution.

To broaden the acceptance and the use of these simple analytic mechanisms to annotate linguistic information, a standardization process via the ISO group on linguistic annotation (TC37/SC4WG6) will be initiated. This initiative might result in a new annotation scheme under the umbrellas of both TEI and ISO and would follow in this respect two other TEI/ISO modules used for linguistic annotation, i.e. "Feature Structures" and "Transcription of Spoken Language".

The primary goals of our contribution are to present the new features to the TEI members and other participants of the conference, and to gather feedback on suggested further developments.

Using the TEI in Training and Research as an Institutional Strategy towards a DH Competence Center

Anne Baillot[5], François Vignale[5]

The development of digital-based research in the context of French universities is favored by the existence of both national (huma-num, HAL) and European (DARIAH, CLARIN) infrastructures supporting archiving, access, good practices and connections between digital research projects. Still, each University can weigh in to define how to develop digital-based research in the Arts and Humanities, and develop their own institutional strategy.

[1] Institute for the German Language, Germany
[2] Berlin-Brandenburg Academy of Sciences and Humanities, Germany
[3] Northwestern University, USA
[4] University of Cologne, Heidelberg University, IDS Mannheim, Germany
[5] Le Mans Université, France

In this paper, we would like to show how Le Mans Université has been able to seize the opportunity of emerging digital-driven projects in the Humanities to conceive a TEI-based strategy that brings together research, teaching, training and innovation. The goal of this paper is to show how the decision to have all projects work with a TEI-based data model allows to develop a strong DH profile in the fields of textual studies and cultural heritage research.

In a first step, we will present the way different projects have been brought together in terms of data modeling based on the TEI (Berlin Intellectuals, READ-IT, ECRIPER). In a second step, we will show how we have developed training modules on graduate level to gain a TEI fluency among a solid group of graduate students and PhD candidates to back the research projects. We will present the pivotal role played by the University Library in this concept. The end of the paper will be dedicated to the connecting options beyond the local university, on a regional, national and international level.

"Examples Work More Forcibly on the Mind than Precepts" – Expanding and Improving the Use of Examples by the TEI Guidelines

James Cummings[1]

At the beginning of Henry Fielding's *The History of the adventures of Joseph Andrews and his friend Mr Abraham Adams* it starts with the idea that "It is a trite but true observation, that examples work more forcibly on the mind than precepts"[2]. This is something that all who have undertaken to teach the TEI Guidelines will recognise, that students more easily latch onto examples of text encoding than explanations of the rules that might govern them. Using this work from Fielding as a starting point, and indeed its use in the TEI Guidelines as an example in itself, I will survey the way in which the TEI Guidelines and Customisations implement examples.

There are a number of shortcomings with the current system and I will propose improvements that ameliorate these. At the core is the suggestion that examples should be stored separately from the Guidelines prose and pointed to from the prose as needed, being virtually included when the Guidelines, or particular customisations, are generated. This separation of the examples from the prose has many benefits including the ability for a single example pointed to from multiple locations (rather than having separate copies as now), or that each of these examples could form an internationalised corpus of fully-valid sample files demonstrating TEI encoding (with their own TEI Header metadata) for students of TEI. As the Guidelines would point into these files for examples different parts or granularity of a single source could be used as examples for different elements. Moving to such a system would not preclude use of the current system and would give a variety of benefits that deserve exploration.

[1] Newcastle University, UK

[2] See http://www.gutenberg.org/files/9611/9611-h/9611-h.htm#book1chapter1

[Day 3 (Tue) 15:30-17:00] SP Session II (Room A4)

Historical Social Networks in Chinese Buddhism

Marcus Bingenheimer[1]

This short paper presents first results of an ongoing project that aims to establish a dataset for the *historical social network analysis* for the study of Chinese Buddhist history. There are two main sources for the data: Buddhist biographical literature on eminent monks (*gaoseng zhuan* 高僧傳) and lineage data connecting masters and students. The former is especially rich for the time between 300 and 1000, when the major gaoseng zhuan collections allow us to situate people in place and time and trace their relationships. The lineage data is extracted mainly from the literature of the Chan school (collected sayings (*yulu* 語錄), lamp-transmission (*denglu* 燈錄)), temple gazetteers (*shanzhi* 山志, *sizhi* 寺志), and other forms of Buddhist historiography.

The data is made openly available on a bi-annual basis and is distributed and archived online, while annual workshops are conducted to train graduate students in using data and tools. Our dataset is so far the most comprehensive tool to view Chinese Buddhist history in network terms. Zooming into regions of the network historians can research the relationships between players at a time, as each relationship-link between actors is referenced to canonical and para-canonical sources. We are now exploring how traditional network measures such as different forms of centrality, or techniques such as clique identification can be put to use to improve our understanding of Buddhist history in East Asia. Research questions include: How can different centrality measures help to guide a historian's search for who is "important" in Chinese history? How bad is the gender gap in Buddhist history and how can be visualized in a network? What connections are there between the structure of the sources and the shape of the network?

The current dataset is available at: http://mbingenheimer.net/tools/socnet/index.html

Complex Manuscript Texts as Prototypes for the Construction of a Source-edition Environment: The Case of Philosophical Glosses

Emmanuelle Kuhry[2]

The Oxford gloss was copied around the most ancient latin translations of Aristotle's treatises on nature. Texts and glosses were gathered into manuscript "text-books" which composed the course material of the students in philosophy at the Faculty of Arts in the 13th century. While coherent enough to allow to identify most glosses from one manuscript to another, the manuscript tradition presents considerable variations, because of the supposedly oral transmission of the content of the glosses, probably copied by the students after hearing lectures. Moreover, glosses and main text often have a distinct manuscript tradition.

Until now, these characteristics and the space constraints of paper, as well as copyright issues have prevented scholars to establish a complete critical edition including the glosses and the main text.

[1] Temple University, USA
[2] Institut de Recherche et d'Histoire des Textes, CNRS, France

Furthermore, the gloss is made of a patchwork of quotations. Identifying these allows to underline the intertextuality between the gloss and medieval literature, and to establish a relative chronology of the circulation of the gloss. It is a true observatory of the medieval teaching methods.

The TEI encoding opens brand new perspectives for those kinds of "multi-dimensional" texts. An open source and collaborative framework will take advantage of this experience to offer tools and methodologies for the encoding of sources in XML-TEI. This paper will present some of the issues encountered and the solutions adopted. In particular, glosses and main text are encoded in separate files, allowing to build a gloss index, which makes it easier to manage the "moving" aspect of the glosses, as they can be bound to a word or another depending on the manuscript. This structure, associated with the use of the "Critical Apparatus" module (parallel segmentation method) allows to reconstruct each manuscript's corpus of glosses, and to manage all the variant readings together, which makes the philological analysis easier.

Francophone Diaries in the 19th Century Russia: A TEI Encoded Edition and Corpus

Alexei Lavrentiev[1], Michèle Debrenne[2], Nina Panina[2], Dmitry Dolgushin[2], Andrey Borodikhin[3]

In this paper we present a project of digital edition and annotated corpus of two diaries written in French by Russian aristocrats in the 1st half of the 19th century. The first one was written in 1812-1813 during the war against Napoleon by a young officer Alexandre Chicherin (1793 – 1813). Its peculiarity consists in a great number of author's drawings that closely interact with the text and require special encoding for further analysis and annotation. The second diary was kept by Olga Orlova-Davydova (1813 – 1876) for almost 20 years (from 1830 to 1847). It contains valuable details on the everyday life of that period and mentions a certain number of historical figures that the author knew personally. Both diaries have considerable value for historical and literary studies but they are particularly interesting for research on the interaction on Russian and French languages and cultures. Therefore, special attention was paid to the encoding of code-switching (inclusion of Russian words and phrases in the French text) and of various kinds of errors or deviations from standard French of the 19th century.

The workflow of the project consists in primary transcription with Microsoft Word using styles and some project specific micro-syntax (special characters) to pre-tag as much useful information as possible. This choice was made due to the project context, as the members of the Russian team have no opportunity to attend a TEI training and cannot afford purchasing a user friendly XML editing software, such as Oxygen XML Author. The approach is interesting, as it allows testing to what extent complex TEI markup can be prepared using a conventional text processing software. The basic text structure (divisions, paragraphs) and some common tags (such as <persName> or <add>) are automatically converted to TEI by OxGarage. In more complex cases, such as <choice> with various sub-elements, or nesting markup (a word in a foreign language that contains an error) we had to use custom style names (e.g. "persName-lang-ru") or special characters (e.g. square brackets for additions and deletions inside a word: pren[aie/a]nt).

The Word transcriptions are converted to TEI XML using OxGarage and the imported to TXM through a series of XSLT transformations. These transformations allow converting custom styles (rendered as <hi> with @rend bearing the style name by OxGarage) and special characters to proper TEI markup, and to prepare tokenization,

[1] CNRS, France
[2] Novosibirsk State University, Russian Federation
[3] State Public Scientific and Technical Library of the Siberian Branch of Russian Academy of Sciences, Russian Federation

indexing and rendering of the text by TXM. TXM provides therefore both an interface for reading, searching and analysing the texts TXM and TEI conformant versions of source documents. These can be used for further annotation and for integration to other publishing and corpus analysis platforms. The prototype of the edition of both diaries is available at the TXM demo portal:
http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones.

The same workflow (with some variations) has already been used in other projects including the edition of Guido Parato's health treaty and integrating Métopes and TXM (both projects are presented at this conference).

Guido Parato's Health Treaty: A TEI Edition on the TXM Platform

Alexei Lavrentiev[1], Elena Markova[2]

Guido Parato was a personal physician of Francesco Sforza (1401-1466), the Duke of Milan. In 1459 he wrote a book on health-keeping that he dedicated to the Duke of Burgundy. The original Latin text was very quickly translated into French, and one of the two extant manuscripts of the French translation is kept in the Russian National Library in Saint-Petersburg (Fr.Q.v.VI.1). This is the base manuscript of our edition.

Parato's work has never been edited. It is not quite original: treaties on health-keeping (or *régimes de santé*) were numerous in the 15th century, and most of the recommendations and remedies it contains can be found in many of them. Nevertheless, it is a good example of "scientific popularization" of its time and deserves consideration in the corpus of Renaissance medical literature.

Our digital edition project relies on the workflow elaborated during the edition of *Queste del saint Graal* (http://catalog.bfm-corpus.org/qgraal_cm). It includes primary editing with Microsoft Word using some styles and special characters. The documents is converted to TEI XML using Oxgarage and then processed by a series of XSLT stylesheets that generate a fully TEI conformant document. The TEI document is finally processed by TXM using "XTZ" import module (http://textometrie.org) and published online on the *Base de français médiéval* portal (http://txm.bfm-corpus.org/?command=documentation&path=/SANTE). The conversion process is fully automated, however once the TEI document is generated, it becomes the base for further editing and annotation. In the presentation we will provide more details on the encoding choices and on the workflow organization.

Publication and Usage of TEI Data in UTokyo Digital Archives Development Project

Satoru Nakamura[3]

The University of Tokyo, mainly University of Tokyo Library System and Information Technology Center, are promoting the "UTokyo Digital Archives Development Project" to make the university's academic assets public and reusable. This research explains the activities of this project as "data provider" providing TEI data and the case study as "data user".

As a "data provider", we provide a digital archive system using Omeka, and publish images and metadata collected from departments of our university. We actively introduce international standards and information technology to support the use of

[1] CNRS, France
[2] Belgorod State University, Russian Federation
[3] The University of Tokyo, Japan

academic assets by the third parties and computers. Specifically, the metadata of the collected materials is provided as LOD, and the images are provided with IIIF. Text data provided by departments is published as TEI data. Since the incoming data is basically provided in plain text, we mark up them with the second encoding level in the guideline “TEI Library Best Practices” and publish them with TAPAS. Currently, 8 collections have been published and 5 materials including in preparation have been associated with TEI. By publishing data conforming to standards such as TEI, we support the utilization of data and promote the development of the systems to enhance the continuity of the project.

To demonstrate the benefits of publishing TEI data, we developed several applications using data published in our project. We used TEI data consisting of 72 court records of the Muromachi shogunate. We divided paragraphs of each case record and normalize the litigation date. Using this structured data, we created an application that visualizes the number of records per month. Furthermore, an application that can display images and text simultaneously was developed, by converting marked-up data to the table of contents information, ranges in IIIF manifest.

We aim to build a system that leads to enhancement of provided data based on the results of researchers and collaborators. On the other hand, there is a limit for librarians to mark up text data with TEI, due to lack of expert knowledge of the materials. Therefore, in order to realize cooperation between researchers and library, it is required to provide systems which give incentives for researchers to create and publish basic research data such as TEI data.

[Day 3 (Tue) 17:15-18:45] LP Session III (Room A3)
Modeling and Metadata

TEI Reclassified

Hugh Cayless[1]

TEI employs two types of class in its definition, attribute classes, membership in which dictates the attribute content of an element, and model classes, which are used to group elements and dictate where they may appear.

A look at TEI's model classes quickly reveals that they fall into essentially two camps, the first groups elements according to their possible location (with names like [element name]Part), and the second according to similarity ([element name]Like). The latter names at least imply some affinity among members, but a closer examination reveals all sorts of problems, such as elements with only loose semantic connections being bundled together and similar elements not grouped. This kind of sloppy grouping accounts for a lot of the incidental complexity in the creation of TEI documents. Moreover, despite its sometimes-semantic associations, the class system is unable to represent some of the semantic features of TEI. The <persName> element, for example, is clearly a type of <name>, which is a type of <rs>. Instead of this relationship being made explicit, however, all are flatly grouped under model.nameLike.

Clearly (and understandably), the function of the class system in TEI is practical: it simplifies the content models of TEI elements by permitting reference to named groups of elements instead of to individuals. Its form and function is in fact a slightly more subtle example of the Durand Conundrum (see Burnard 2013, http://journals.openedition.org/jtei/842). It maps exactly onto the concept of expandable <define> in Relax NG (http://relaxng.org/spec-20011203.html#define-ref). In the same way that the introduction of Pure ODD (see http://www.tei-c.org/release/doc/tei-p5-doc/readme-3.0.0.html) has enabled the TEI to peel itself away from strict reliance on Relax NG's syntax, a new class system could enable a clearer expression of TEI's semantics and allow for less incidental complexity, easier extraction of information content, and improved interoperability.

A refactoring of the TEI class system could be accomplished without any changes to the basic TEI infrastructure, by renaming and adding classes and changing element class memberships. Going further would involve changes to the way TEI works. It is clear from looking at the grouping patterns in http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-model.nameLike.html, for example, that inheritance would be a useful feature. Elements themselves could define classes, so <name> could inherit directly from <rs>, and persName from <name>, rather than relying on a separate class system for associating them. A thorough investigation of the *Part models might reveal that they could be refactored into semantic classes without unnecessarily complicating the content models of the elements that can contain them.

Ideally, A refactored class system for TEI should also do away with the confusion between the ontological status of an element and where it can occur. <lang>, for example, is arguably a sort of name, and is thus a member of model.nameLike, but the Guidelines treat it only as a member of <etym>. It gets into the latter's content model via model.phrase(!) and because of its classification, is available in very many contexts for which it was not designed, and for which there are no examples in the Guidelines. A possible solution would be to have two ways of indicating class membership, one which results in the derived element inheriting content model membership from the parent class

[1] Duke University, USA

and one which does not. A more direct one would be to avoid grouping elements just because they have similar meanings.

A complete replacement for TEI's current class system would be a large change, perhaps warranting a new revision (P6 from the current P5). This paper will outline some possible solutions to the problems the current system suffers from with a view toward gathering community feedback and developing a roadmap for "reclassifying" TEI.

Modelling Authentication

Sean Michael Winslow[1]

This paper will discuss issues in the modelling of authentication across time periods and regional boundaries, from the existing <seal> element to medieval European chirographs to Japanese *Jitsu-in* to ways to model authenticating factors of emails and other digital objects. Such authenticating elements as seals and stamps can be considered paratexts, as they provide information which changes the interpretation of the associated text. Accordingly, content-based markup like the TEI should have a way to capture authenticating elements in order to present the necessary context for the interpretation of the text. This paper will introduce the need for such modelling from a theoretical standpoint and present a working ontology of methods of authentication, how they have been modelled in the CEI, and how it is being implemented in the ODD of the current project (in progress).

The proximate need for such an extension is the mapping of documents marked up by the Charters Encoding Initiative to TEI P5. Nevertheless, the practices modelled are similar across time periods and regions, from ancient Mesopotamia to medieval England to modern Japan. The paper will consider not just charters and documents, but also text-bearing museum objects that have authentication features, such as banknotes and passports and born-digital objects such as e-mail. All of these items bear text, but that text is only meaningful in the context of their authenticating features, suggesting that the TEI needs to expand its concept of what is important to mark up.

This work is part of the proceeds of the FWF-Projekt "Retain Domain Specific Functionalities in a Generic Repository with Humanities Data" (PI: Georg Vogeler), itself arising from the the preservation of documents in the "Illuminierte Urkunden als Gesamstkunstwerk" project.

References

Boyle, L. (1992) "Diplomatics." in Powell, J. *Medieval Studies: An Introductio*n, *Second Edition*. Syracuse University.

Cummings, et al. "<seal> element description should be expanded to other authentication mechanisms #1655" https://github.com/TEIC/TEI/issues/1655

Duranti, L. "InterPARES Authenticity Task Force Report" http://www.interpares.org/book/interpares_book_d_part1.pdf

Maria Milagros Cárcel Ortí, ed. (1997) *Vocabulaire international de la diplomatique*, 2. ed., Valéncia: Collecció Oberta.

Park, E. (2001) 'Understanding "Authenticity" in Records and Information Management: Analyzing Practitioner Constructs'. The American Archivist: Fall/Winter, Vol. 64, No. 2.

Vogeler, G. "Charters Encoding Initiative." http://www.cei.lmu.de/index.php

Vogeler, G. (2018) "Digital Diplomatics: The Evolution of a European Tradition or a Generic Concept?" in Cubelic, et al. (eds..): *Studies in Historical Documents from Nepal and India* (Documenta Nepalica – Book Series, Band 1). Heidelberg: Heidelberg University Publishing, 85-109.

[1] Austrian Centre for Digital Humanities, Austria

Tei-Meta: A Tool for Editing Metadata in TEI - Application to Oral Language Research Purposes

Christophe Parisse[1], Carole Etienne[2], Loïc Liégeois[3]

The TEI is a perfect standard for storing and preserving digital texts whatever their nature and level of granularity. It can also represent spoken language data based on a textual transcription (ISO-TEI[4]). More recently, we got interested by a common solution to describe all the spoken corpora for a research purpose. This means that a researcher who knows nothing about a corpus should understand if this corpus is interesting or not for his own studies, just reading the metadata. This will make easier to reuse already existing material with good metadata. For this purpose, TEI appears as a good solution regarding its semantic organization and the gathering of metadata and data in a same file. To make this possible, we have been working in two parallel directions:

1. Creation of a reference for spoken language metadata. This reference is much more powerful than the often used Dublin Core/Olac information. In particular, we selected elements and properties available in the TEI header to create a reference for fine-grained information about speakers, settings, recording quality, and languages used. This reference does not try to be exhaustive. On the contrary, we based our work on the experience from the Orfeo project[5] on already existing corpora so as to devise a middle ground between too much and too few information. What we suggest is a minimal set of (usual) information relevant for research purpose (selection, documentation, helpful during analyses) with modular add-ons if necessary.
2. Creation of an open-source web-based multilingual tool (available also as a standalone application) that can edit any type of TEI header file, including spoken language metadata. The tool main interest is that it is able to work with any TEI ODD file that contains the description of the metadata to be edited, each ODD file generating its own representation. It is thus easy to create as many version of the tool that they are ODD versions, regardless of language. Stylesheets or ODD information can be added for specific visual presentation (on the base of CSS format). The tool and the ODD files are described and can be downloaded on the web (http://ct3.ortolang.fr/teimeta/readme.php). The sources are available on GitHub and an online recommended version is available at http://ct3.ortolang.fr/teimeta.

[1] INSERM/Modyco, U. Paris Nanterre, CNRS, France
[2] ICAR, CNRS, France
[3] LLF/CILLAC-ARP, U. Paris Diderot, France

[4] ISO-TEI - European group working on "Transcription of Spoken Language" - ISO 24624:2016 - Language resource management -- Transcription of spoken language. (s. d.).

[5] Benzitoun C., Debaisieux, J.-M., Deulofeu, J. Le projet ORFÉO : un corpus d'étude pour le français contemporain, Corpus 15, Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

[Day 3 (Tue) 17:15-18:45] Panel Session I (Room A4)

Confronting Challenges in Marking Up Pre-modern East Asian Documents

Chair/Discussants: **Marcus Bingenheimer[1], Duncan Paterson[2], Martin Holmes[3]**
Organizer: **Hanna McGaughey[4]**

Panel Introduction

The TEI desire for broad applicability is repeatedly frustrated by the unique needs and preferences of East Asian text encoders. The TEI Guidelines were initially developed by committees of roughly equal Europe and North America representation to markup "any document written in any language during the entire span of history." However, encoders of East Asian texts confront difficulty finding appropriate TEI tags, accommodating varying text models, and managing and sharing good encoding practice. Nonetheless, these encoders turn to TEI because of its high level of standardization.

To expand TEI applicability to East Asian texts, we will share our experience both in encoding manuscript and print texts produced before significant contact with European text traditions and in revising guidelines as a member of the Technical Council. Drawing on work with pre-modern Japanese performance treatises, Hanna describes complexities of texts written in a combination of three types of characters. Duncan points out challenges in the practical application of the guidelines to markup textual features of early Chinese woodblock print encyclopedias. Martin argues that TEI needs a new working group to focus on CJK languages of all time periods.

Pre-modern Japanese Texts and the TEI Guidelines: A Report from the Frontline of Encoding Zeami's Treatises

Presenter: **Hanna McGaughey**

Japanese may be the most complex writing system in the world, with its three different components: Chinese characters and, derived from them, two phonetic lettering systems (Jp. *katakana and hiragana*). All three scripts appear in mixed kanji and kana texts (Jp. *majiri bun*) like modern standard Japanese, but are found also in pre-modern documents. Japanese writers have long used markup to accommodate script variety. Indeed, katakana were created as a shorthand pronunciation guide to gloss Chinese sutra texts for Japanese monks' recitations. Such use of interlineal glosses (Jp. *furigana*) left the prayer halls and soon provided a greater variety of semantic as well as phonetic props to readers of all kinds of texts. Additional features such as pre-modern punctuation (Jp. *kudokuten*) marking Japanese word order in Chinese and Sino-Japanese texts and two smaller lines of text in the space of one line (Jp. *warigaki*), not to mention markup and annotation found in modern critical editions, guide today's readers of pre-modern Japanese texts.

All of the textual features I mentioned here and more may be found in treatises by actor, troupe leader, playwright, and composer Zeami (ca. 1363 - ca. 1443) about the performing arts, including pedagogy, aesthetics, and performer-audience communication. In my efforts to produce a machine-readable version of these texts for electronic analyses of difficult-to-define terms, TEI recommends itself as a highly standardized method of marking up complex text features.

[1] Temple University, USA
[2] Heidelberg University, Germany
[3] University of Victoria, Canada
[4] University of Trier, Germany

As my translations above suggest, Japanese markup strategies correlate approximately to strategies developed in European and North American document production. Nevertheless, the application of TEI Guidelines in marking up Japanese texts poses problems concerning appropriate element tags, good encoding practices, and so forth. In this paper, I will elaborate on the challenges of marking up pre-modern Japanese texts by drawing on examples from my efforts to digitize Zeami's treatises.

I Just Want to be Normal: Post-Colonial Reflections on Encoding Early Modern Xylographs

Presenter: **Duncan Paterson**

TEI's goal to "apply to texts in any natural language, of any date, […] without restriction" often raises suspicion among area specialists, where post-colonial critique has repeatedly deconstructed similar universalist claims. Indeed, while the oldest known printed text and the quantitative bulk of pre-modern books uses literary Chinese, its conventions and terminology remains absent from the Guidelines.

My paper addresses three kinds of challenges that this lacuna presents to encoders of East Asian sources. Beginning with conceptual biases that unduly project European codicological conventions ('edition', 'font', 'hand', etc.) onto non-Western sources. Using examples from xylographic prints from Early Modern China, I will discuss the evolution of emphasis markers in manuscripts (Ch. *jù dòu* 句讀) and print (Ch. *kàn zhòng hào* 看重號), to demonstrate the logical challenges that rubicated glyphs present to TEI's textual model. I conclude with a discussion of means for overcoming technical biases inherent in XML, such as the flow attribute, or the role of different whitespace characters in marking up text without word boundary markers. TEI can play a crucial role vis-à-vis competing standards that currently hinder the interchangeability of historical CJK documents.

A Coordinated Approach to Improving Support for CJK Languages in the TEI Guidelines

Presenter: **Martin Holmes**

As the use of TEI for historical and literary CJK texts grows, the need for enhancements and encoding recommendations for textual features not seen in the western traditions is increasingly obvious. In this paper I will argue that, rather than deal with these issues on a ticket-by-ticket basis, we need a working group to focus on CJK languages, bringing together expertise from different languages and periods, to propose a set of harmonized changes to the Guidelines.

In 2013, a working group contributed new sections of the TEI Guidelines related to character/glyph orientation and text-directionality. Much of the work of this group involved determining where the boundaries lie between what the TEI should handle and what can be left to other standards such as Unicode, BCP 47 , and Cascading Stylesheets. Where existing standards appear to provide effective mechanisms for encoding, we may defer to these standards, but there is still a requirement to show how they can be deployed within TEI markup, and if possible recommend a single model for usage, in the interests of interoperability. Where existing standards do not meet requirements, new elements, attributes and values, along with recommendations for their use, must be added. However, where external standards are suitable but not quite sufficient for a specific need, I would like to see the working group engage directly with external standards bodies to make recommendations, in preference to implementing TEI workarounds. For example, should we adopt HTML5's ruby elements for encoding of features such as *kudokuten*, and if so, do we need to propose enhancements to HTML5 and CSS to the W3C to handle historical features not covered by the current standards? Such engagement takes time and patience if it is to be fruitful, but the long-term gains in interoperability can be substantial.

References

Internet Engineering Task Force (2009). *Tags for Identifying Languages (BCP 47; RFC 5646).* Accessed April 6. 2018. https://tools.ietf.org/html/bcp47.

Kawase, Akihiro, Taro Ichimura, and Toshinobu Ogiso (2013). "TEI:P5 ni motozuku kinsei kōgoshiryō no kōzō to sono mondaiten." *Shimonkon* 2013 *Ronbunshū* 2013, no. 4: 7–12.

Liú Zǐmíng 劉子明, and Ān Zhèngtáng 安正堂 (1612). *Xīnbǎn Zēngbǔ Tiānxià Biànyòng Wénlín Miàojǐn Wànbǎo Quánshū* 新板增補天下便用文林妙錦萬寶全書. Vol. 38. China: Shūlín Ān Zhèngtáng, Liú Shuāngsōng. 書林安正堂, 劉雙松.

McDermott, Joseph Peter (2006). *A Social History of the Chinese Book: Books and Literati Culture in Late Imperial China, Understanding China*. Hong Kong: Hong Kong University Press.

Said, Edward (2003). *Orientalism*. 25th Anniversary Edition. London: Penguin Books, Limited.

Sperberg-McQueen, C. M., and Lou Burnard (1995). "The Design of the TEI Encoding Scheme." *Computers and the Humanities* 29, no. 1: 17–39. https://doi.org/10.1007/BF01830314.

"TEI: East Asian/Japanese SIG." Accessed April 6, 2018. http://www.tei-c.org/Activities/SIG/EastAsian/.

"Text Directionality Workgroup - TEIWiki." Accessed April 6, 2018. https://wiki.tei-c.org/index.php?title=Text_Directionality_Workgroup.

Wood, Frances, and Mark Barnard (2010). *The Diamond Sutra: The Story of the World's Earliest Dated Printed Book*. London: British Library Board.

[Day 5 (Thu) 9:00-10:30] LP Sessions IV (Room A1)
Proposals and Recommendations

Using ODD for HTML

Martin David Holmes[1]

Although the ODD (One Document Does It All) language is normally used for the creation of TEI customizations or extensions, it is a highly-effective tool for editors working in other XML markup languages. This paper will discuss the use of ODD to define a highly-constrained schema for HTML5, to enforce stylistic and encoding practices, define custom attributes and value-lists, and enable easier editing and validation of project content in the Oxygen XML Editor environment. I will provide a brief history of the project, whose first incarnation, created with the DreamWeaver HTML editor, was somewhat chaotic from a code point of view, and show how the implementation of an ODD-based schema provides huge advantages for authors, editors and encoders, as well as substantially simplifying the code itself.

1. Introduction

Although ODD (One Document Does it All) is a feature of the TEI language, and used primarily for creating TEI schemas, ODD in fact "goes beyond this to provide a generic tool for the documentation and management of any XML encoding scheme, not necessarily one based on the TEI" (Burnard and Rahtz 2014). Syd Bauman 2017 points out that the TEI ODD language "can be used for two related, but distinctly different purposes: 1) to *create* a markup language, including documentation and schemas; and 2) to *customize* a markup language that was already written in ODD." This paper describes a use-case which does not quite fall into either of those categories: the use of ODD to create a highly-constrained customization of a markup language not originally written in ODD. The language in this case is HTML5, in its XHTML incarnation.

In 2017, our unit was approached by a faculty member who had for many years been building a fascinating website called "Mapping Keats's Progress"; it is in one aspect a biography of the poet John Keats, but it has many other features. The site had been developed by the author and a collaborator using the DreamWeaver web-authoring software. Although its interface and structure were functional and attractive, the HTML code had become a huge mass of incomprehensible nested structures, including 24 JavaScript and 69 CSS files. An example of the kind of needless complexity that had resulted from the dependence on a WYSIWYG tool to manage style and layout can be seen in Figure 1. In the fall of 2017, I began the process of rewriting it, with the aim of keeping it as simple as possible while reproducing and enhancing the design and functionality. The result has only one CSS file and less than 100 lines of JavaScript, none of which is essential.

[1] University of Victoria, Canada

```
<div class="p7TP3cwrapper_03">
  <div id="p7TP3cvp_1" class="p7TP3_vp">
    <div id="p7TP3pw_1" class="p7TP3_slide_wrapper">
      <div id="p7TP3w1_1" class="p7TP3_panel">
        <div id="p7TP3c1_1" class="p7TP3_content_03 p7TP3content">
          <div class="p7tp3-col-wrapper no-columns"> </div>
        </div>
      </div>
      <div id="p7TP3w1_2" class="p7TP3_panel">
        <div id="p7TP3c1_2" class="p7TP3_content_03 p7TP3content">
          <div class="p7tp3-col-wrapper no-columns">
            <div id="p7AP3_1" class="p7AP3-02 p7AP3 responsive">
              <div id="p7AP3tb_1" class="ap3-toolbar closed"><a href="#" title="Hide/Show
                Menu">&equiv;</a></div>
              <div id="p7AP3rw_1" class="p7AP3root-wrapper closed">
                <div class="p7AP3trig p7ap3-theme-02">
                  <h3><a href="#p7AP3c1_1" id="p7AP3t1_1">16 March 1816: Keats, Joseph Severn,
                    and Chivalric Infatuations </a></h3>
                </div>
                <div id="p7AP3w1_1" class="p7AP3cwrapper p7ap3-theme-02">
                  <div id="p7AP3c1_1" class="p7AP3content p7ap3-theme-02">
                    <div id="p7AP3p1_1" class="p7AP3panelcontent p7ap3-theme-02">
                      <div class="p7ap3-col-wrapper multi-columns">
                        <div class="p7ap3-column width-50">
                          <div class="p7ap3-column-content p7ehc-2">
                            <h1>6 Goswell Street Road, London</h1>
```

Figure 1. The long journey to the <h1> element in the original DreamWeaver-generated HTML.

2. Why not use TEI?

Although the original HTML had got severely out of control, the original content was already basically complete, and encoded in HTML. I was able to use a toolchain consisting of HTML Tidy, XSLT and Python to clean up and simplify the content to a point where it required only some proofing and enhancement, and since the project author was already familiar with HTML, but not with TEI, and the markup itself was relatively simple, it seemed easier to stick with HTML5 encoding. Given a sufficiently rigorous schema for that encoding, it would be trivial to generate TEI from the markup if we wanted it in the future.

3. Why use ODD?

The W3C provides an excellent validation tool for HTML5 in the form of the Nu Html Checker. This is the tool we use for final validation of all HTML sites we produce. However, it is of course a generic tool; it checks conformance against the entire schema[1]. I wanted to constrain the HTML quite aggressively, to define custom attributes (as HTML5 allows) with closed value-lists, and to incorporate Schematron rules, to ensure that the site style and structure remains consistent throughout the document set. We also require documentation of the rules, along with encoding guidelines and examples. ODD is the perfect choice for this.

This paper will discuss the use of ODD in this project, providing examples of custom attributes, Schematron and other features, and show how we keep the encoded scholarly text separate from the global site boilerplate content (menus, headers, footers and so on), uniting them only at build time to publish the site.

Keywords

ODD, HTML, non-TEI projects

References

Bauman, Syd (2017) “tei_customization: A TEI customization for writing TEI customizations.” Paper delivered at TEI 2017 Conference, Victoria, B.C., Canada, November 14. https://hcmc.uvic.ca/tei2017/abstracts/t_110_bauman_teicustomization.html.

Burnard, Lou and Sebastian Rahtz (2014) “ODD: One Document Does it All.” Workshop delivered at TEI 2014 Conference, Evanston, Illinois. http://www.tei-c.org/Vault/MembersMeetings/2014/workshops/odd-one-document-does-it-all/index.html.

[1] In fact, any of multiple schemas, depending on the input document type.

From File Interoperability to Service Interoperability: the Distributed Text Services

Bridget May Almas[1], Hugh Cayless[2], Thibault Clérice[3], Vincent Jolivet[3], Emmanuelle Morlock[4], Jonathan Robie[5], James Tauber[6], Jeffrey C Witt[7], Pietro Liuzzo[8]

The Canonical Text Services (CTS) protocol[9] has allowed many classical, canonical texts encoded in TEI to be made available in a machine-actionable, linked open data fashion. However, CTS is tightly coupled to its text identifier system, which is not adaptable to the citation systems used by more modern content or other forms of writing, such as papyri or inscriptions. CTS also suffers from API design issues, scalability of response formats, and poor definition of response MIME types. These issues have limited its application outside of the set of classical texts for which it was originally designed.

To address these limitations, a group of interested scholars and technologists have created a simpler design for textual data sharing that follows best practices for REST APIs. The Distributed Text Services (DTS) specification has been designed with three major API endpoints that each have their own role: catalog browsing (Collection), reference browsing (Navigation) and text retrieval (Document).

The DTS API is based on a standard for Hypermedia-Driven Web APIs called Hydra[10], using JSON-LD and Dublin Core (DC) as core metadata vocabularies. Reuse of these standards should facilitate adoption and uptake. The Collection endpoint offers navigation of text collections; the Navigation endpoint offers querying for the specific references within a text or part of a text; the Document endpoint offers query and retrieval of complete or partial texts. The first two return JSON responses defined by the Hydra standard, the third returns the TEI XML of the requested text or fragment.

The paper will discuss the origins of this effort, the challenges that the data providers faced using other protocols, the specific use cases, the final API structure, and the decision making process. We hope DTS will be of interest to the TEI community at large, as a general purpose TEI discovery and retrieval API.

[1] The Alpheios Project, Ltd
[2] Duke University, USA
[3] École Nationale des Chartes, France
[4] HiSoMA Research Center, France
[5] biblicalhumanities.org
[6] Eldarion
[7] Loyola University, USA
[8] Universitat Hamburg, Germany
[9] http://cite-architecture.github.io/cts/
[10] https://www.w3.org/community/hydra/

TEI-Lex0 Etym – Towards Terse Recommendations for the Encoding of Etymological Information

Jack Bowers[1,2], Axel Herold[2,3], Laurent Romary[3,4]

The current TEI guidelines only offer very loose constructs for the representation of etymological information in TEI encoded dictionaries. With the generic free text <etym> element, possibly combined with inline elements such as <lang> or <mentioned>, the guidelines offer too much flexibility and a lack of guidance to reflect precise linguistically-founded analyses. In a context where more and more lexical resources are being created and legacy print (including etymological) dictionaries are available in digital form, providing better interoperability between these resources is essential to enabling precise cross-resource analyses on diachronic phenomena across languages.

Three initiatives have created favourable conditions for this work:

- The COST action ENEL[5] (ended September 2017) which initiated a series of workshops to identify general constraints on the encoding of dictionaries (cf. Romary & Tasovac, submitted) named TEI-Lex0; [+ follow-up in Elexis]
- Reserialization of the LMF standard (ISO 24613) which will include an etymological extension (ISO 24613-3) and a TEI serialization (ISO 24613-4)
- The creation of a DARIAH Working Group on lexical resources which is sponsoring two TEI-Lex0 workshops in 2018

Building off of recent efforts adressing etymology in TEI (cf. Bowers & Romary, 2016), TEI-Lex0 Etym defines a more restrained set of options for encoding any given single phenomena which are designed to be able to equally handle born-digital and retro-digitized print sources. The scope of our proposal covers the usage of the following concepts central to etymological description:

- Etymology element (<etym>): structuring etymology processes through typing and recursivity
- Typology of etymological processes
- Etymons and their forms
- Related forms (cognates, and others)
- Temporality of etymological processes
- (possibly shallow) Bibliographical references in etymologies
- `Prose description of etymological process and content

In our discussion we present examples from a set of historical print Germanic[6] etymological dictionaries as well as from other print and born digital sources.

References

Jack Bowers, Laurent Romary (2017). Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative*, TEI Consortium, 〈10.4000/jtei.1643〉. 〈hal-01296498v2〉

Banski, Piotr, Jack Bowers, and Tomaz Erjavec (2017) "TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms." In *Electronic*

[1] Austrian Center for Digital Humanities (ACDH), Austrian Academy of Sciences (ÖAW), Austria

[2] École Pratique de Hauts Études (ÉPHÉ), France

[3] Berlin Brandenburgisch Akademie der Wissenschaften (BBAW), Germany

[4] Institut National de Recherche en Informatique et Automatique (Inria), France

[5] http://www.cost.eu/COST_Actions/isch/IS1305

[6] https://github.com/xlhrld/retro-dict

Lexicography in the 21st Century: Proceedings of ELex 2017 Conference, 485–494. Lexical Computing. 〈hal-01757108〉

[Day 5 (Thu) 9:00-10:30] LP Session V (Room A2)
TEI as a Tool and Method for Analysis

Shakespeare and the Enumeration of Semantic Universals

Brian L. Pytlik Zillig[1], Mary K. Bolin[1]

Wierzbicka and Goddard (1994) describe a theory of semantic universals, basic meanings common to all human languages. They have tested "Natural Semantic Metalanguage" (NSM) across many languages and have a list of 65 "semantic primes" that can be expressed in English and other languages. The primes include: I, you, someone, be, do, have, think, good, bad, etc. These meanings can only be defined by using more complex semantic units, and are therefore the most basic semantic structure of a language.

One technique of NSM analysis is "explication," in which texts are rewritten using semantic primes. Goddard and Wierzbicka (1994) give this explication of the word "terrible."

> this X is very very bad
> something very bad can happen because of this
> when I think like this, I feel something very bad because of it

TEI texts can be used for the examination of NSM phenomena. NUPOS is a system for analyzing texts grammatically, which can be used with software such as Philip Burns' MorphAdorner, which normalizes and tags words as parts of speech and grammatical role (NUIT 2013). The powerfully granular MorphAdorner analysis and NUPOS categories may be combined with NSM analysis to produce data such as the frequencies of basic meanings like think, know, etc., in early modern drama texts. Further manipulation of the data might yield machine-generated NSM explications of these texts, as in the example above.

We present an exploration of the results of combining MorphAdorner tagging and NSM analysis of 112 TEI-encoded texts by Shakespeare, Marlowe, Jonson, Middleton, and Shirley to compare the presence of particular semantic primes in those texts, including experimental visualizations of the data.

Texts used here mainly originated in the work of the Text Creation Partnership. Martin Mueller provided additional WordHoard texts for the project.

TEI Processing Model - Beyond Books by Dead White Men

Magdalena Turska[2], Wolfgang Meier[2]

TEI Processing Model (TEI PM), recently integrated into TEI Guidelines, is an abstraction layer which provides means of specifying the intended transformations for TEI elements with the TEI ODD metalanguage. Formerly known as Simple Processing Model[3] (an outcome of the TEI Simple project) it was first conceived as a solution for publishing early modern material encoded in TEI or 'printed books by dead white men' as the working

[1] University of Nebraska, USA
[2] eXist Solutions, Germany
[3] TEI Simple https://teic.github.io/TEI-Simple/

summary went. Nevertheless, authors' hopes were that Processing Model can stretch further than limited scope targeted initially.

Since early days of TEI Simple we have periodically reported on continuous development of the TEI PM and the TEI Publisher[1] - software package implementing it in XQuery for eXist-db platform. Experiments showed good results with large corpora of modern texts, including non-Western material[2]. As presented last year, even extending the model for non-TEI vocabularies, e.g. DocBook, was fairly straightforward. Quite early on successful attempt was made to use TEI PM in the context of papyrology[3]. We were thus certain that PM can be employed also to produce high-quality critical edition of richly encoded manuscript material.

Opportunity to prove it arrived with the SSRQ - the Collection of Swiss Law Sources which contains material from the early middle ages until early modern times from all language regions of Switzerland.

This large corpus has been enriched with vast critical apparatus, ranging from transcriptional features, such as additions, deletions or hand shifts to corrections and conjectures, marking of missing or unclear passages, place- and person names along with broad commentary. All this information needs to be accessible for the reader while conforming to long standing customs of presentation followed in the field.

While this project has indeed proved to be challenging, requiring roughly twice as many lines of ODD to specify processing requirements as other editions we used TEI PM so far, the difficulty was rather in tediousness of expressing numerous scenarios for deeply nested encoding via ODD models than on a conceptual level. Furthermore, this exercise of applying TEI PM brought new ideas for improvement of the TEI PM itself. We hope that presenting this use case for the TEI PM will be interesting and beneficial for the TEI community.

Hierarchies Made to Be Broken: A Standoff Approach to the *Frankenstein* Bicentennial Variorum Edition

Elisa Eileen Beshero-Bondar[4], Raffaele Viglianti[5]

This paper addresses the challenges of collating digital editions made at different times by different editors, and it discusses the bicentennial Frankenstein variorum project that the authors are preparing to be released in 2018 in celebration of the bicentennial of Frankenstein's first publication. Comparing documents encoded in conflicting ways encourages a view of TEI XML document hierarchies as "made to be broken," that is, designed to be fragmented into comparable units.

Building on earlier discontinuous digital editions of Frankenstein, we break and transform these source documents into bridgeable unit pieces optimized for machine collation with CollateX[6]. The output of automated collation indicates variation among the documents by converting the markup of the source files into plain text. Finally, we up-convert the simple collateX output to a stand-off format of TEI XML that holds pointers to specific locations in the source editions. This permits us to build an edition interface that preserves the diverse kinds of information encoded in the source editions.

---

[1] TEI Publisher http://teipublisher.com

[2] Cf. Foreign Relations of United States https://history.state.gov/ and SARIT - Search and Retrieval of Indic Texts http://sarit.indology.info/

[3] Corpus of Pyu Inscriptions http://hisoma.huma-num.fr/exist/apps/pyu/index2.html

[4] University of Pittsburgh, USA

[5] University of Maryland, USA

[6] CollateX software applies a graph-based model of text to locate variants in documents. See https://collatex.net/doc/

Stand-off collation identifies variant readings between texts by grouping pointers, as opposed to grouping strings of text according to the parallel segmentation technique described in the TEI Guidelines, Chapter 12[1]. The TEI offers a stand-off method for encoding variants, called “double-end-point-attachment”, in which variants can be encoded separately from the base text. Despite its flexibility, this stand-off collation method requires a “base text” (a single text to be preferred over the other variant texts) to which anchor variant readings. While doing so is a traditional approach to preparing critical editions that long predates the digital age[2], it is not ideal for variorum editions like ours that, by design, do not choose a base text[3]. Our approach, therefore, identifies variance and groups readings without designating a base text and without conflating the source witnesses into one document[4]. The authors share all the stages of their project with documentation of their workflow on GitHub (see https://github.com/PghFrankenstein/Pittsburgh_Frankenstein) and seek to suggest new options for the TEI Guidelines on applications of stand-off annotation in preparing variorum editions.

[1] See especially the TEI P5 Guidelines, 12.2.3 and 12.2.4: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS

[2] See W. W. Greg's "The Rationale of Copy Text", Studies in Bibliography Vol. 3 (1950/1951), 19-36, which famously challenged longstanding conventions of accepting the last text edited by an author as the "best" text.

[3] In 19th-century textual scholarship, for example, William Wordsworth's The Prelude is a textbook example of a poem that altered drastically between its first publication in 1805 and its last in 1851 and there is no single preferred text. Standard print editions of this poem (such as the Penguin edition of 1996) produce as many as four or five versions to be read together in various stages of drafting without designating which is to be preferred. Pertinent to our project, James Rieger's edition of Frankenstein in 1974 presented multiple versions of the novel in a way that challenged the long-standing notion that the last text edited by Mary Shelley (1831) was necessarily the best text, and helped incentivize a new basis for textual scholarship that made the selection of a preferred text an open question rather than a determination of the scholarly editor. We therefore seek to construct a digital edition that does not rely on the selection of a single preferred text. In the the digital arena perhaps our closest "kin' project is Barbara Bordalejo's work on the Darwin Variorum. http://darwin-online.org.uk/Variorum/1861/1861-1-c-1866.html, where it is possible for site visitors to choose for themselves the editions they wish to see and compare, expanding the purview beyond the traditionally best-known work of 1859. The interface we are designing for our Frankenstein Variorum will resemble this one for giving readers the options, and also providing information about each.

[4] For a related example, see Viglianti, R. Music and Words: reconciling libretto and score editions in the digital medium. “Ei, dem alten Herrn zoll’ ich Achtung gern’”, ed. Kristina Richts and Peter Stadler, 2016, 727-755.

[Day 5 (Thu) 9:00-10:30] Panel Session II (Room B1,2,3)

Facilitating the Dissemination of TEI-based Digital Resources in Japan: As Early-Career Researchers from Tokyo Digital History

Naoki Kokaze[1], Soki Oda[1], Boyoung Kim[2], Natsuko Saji[1]

This panel has two aims. Firstly, we are going to introduce some projects which have addressed to disseminate TEI-based data for the humanities studies, among Japanese academic communities. Secondly, we would like to consider what it means to researchers in early career stages that they engage in such DH projects, especially in terms of career-making. In that sense, this panel necessarily introduces pilot projects.

As for the first point, in Japan, there are far less amount of structured open data for humanities research than European and American communities, which encouraging some of the Japanese researchers to take actions. Then, as the first speaker of this panel, Soki Oda is going to present his research utilizing the TEI data provided by the Regesta Imperii Online, as a reference case for us to found TEI-based data. Subsequently, Boyoung Kim introduces projects of digitizing biographical materials held by the Shibusawa Eiichi Memorial Foundation, Natsuko Saji shares with you the challenge of the Iwami Ginzan World Heritage Center, and finally Naoki Kokaze demonstrates a relatively path-breaking project, Engi-shiki.

Consequently, it should be valuable for us to discuss with participants the merits of young researchers' engaging in DH projects, comparing with the cases in the academia of other countries.

Regesta Imperii Online: Visualizing Medieval Papal Legation

Presenter: **Soki ODA**

This paper shows a simple example of a project making use of TEI-encoded texts. The project attempts to visualize the itineraries of papal legates in the Middle Ages by utilizing a dataset of TEI XML documents from Regesta Imperii (RI) Online.

RI chronologically records the activities evidenced by charters and other sources of the medieval German kings and Holy Roman Emperors as well as of the popes and some cardinals in the form of German abstracts. RI Online, in which all printed volumes are available in full text data, provides ca. 130,000 XML files of regesta.

In this paper the itinerary of Dominican friar and cardinal legate Hugo of St. Cher is visualized in TimelineJS, an open-source tool to build interactive timelines. It is difficult to reconstruct medieval legatine journey because of the widespread itinerary and the scattered documents. However, the XML files from RI Online enable us to efficiently analyze and visualize the issue dates, locations, recipients and contents of the legatine documents.

Besides the visualization of the papal legation, I would like to show future prospect of this project. I am planning to collect unedited legatine documents and markup in accordance with the TEI guidelines. The TEI records will be disclosed in TAPAS Project.

[1] The University of Tokyo, Japan
[2] Shibusawa Eiichi Memorial Foundation, Japan

Challenges for TEI-based Representation of Japanese Texts from the Shibusawa Eiichi Memorial Foundation

Presenter: **Boyoung Kim**

In this presentation, I am going to introduce two pilot projects of the Shibusawa Eiichi Memorial Foundation, so as to discuss some issues and challenges in TEI-based representation of the Japanese digital resources.

The Shibusawa Eiichi Memorial Foundation is a non-governmental organization with 130 years of history dating back to 1886. The Foundation is devoted to honoring and promoting the achievements and ideas of Shibusawa Eiichi and the modernization of Japan, especially business history, between the 19th and early 20th centuries. The Foundation holds rich collections of a wide range of historical materials including books, documents, audiovisual materials, artifacts, and even two buildings.

Biographical materials on Eiichi, one of the most important collections of the Foundation, had been accumulated in the sixty-eight volumes of the Shibusawa Eiichi Denki Shiryo. The full collection of the Denki Shiryo is taking up over three shelves of a bookshelf and totaling more than 48,000 pages. It has been digitized since 2004 and about 40,000 pages of texts in digital form with its image files have been available online for free since 2016.

Now the Foundation is considering implementation of the Text Encoding Initiative (TEI) for better preservation and sharing of these huge amounts of texts. Especially, the Ebiki (image index) Database and the Jikken Rongo Shoseidan (series of articles based on Eiichi's talk related to Analects and his experiences) Digital Edition, started as pilot projects in 2018, are going to be the target sources of the TEI implementation. I am going to share with the participants the challenges and problems related to these projects.

Fig 1. Ebiki Database of the Shibusawa Eiichi Memorial Foundation

Liberate Multi-Language Primary Sources using TEI: A Case Study of Iwami-Silver-Mine

Presenter: **Natsuko Saji**

This research project aims to introduce a pilot project: to 'TEI-nize' abundant existing primary sources concerning to Iwami-silver-mine: to liberate the sources to people who do not recognize Japanese language.

Iwami was registered as a UNESCO world heritage in 2007, as it had been the one of the most important silver-mine in the early modern period. The sixteenth-century would be called as 'the century of silver'; silver produced in Japan and South America circulated all over the world, which could encourage to proceed the globalization. It is important to scrutinize regional studies on each mine and to consider comparisons and interactions between each other, when to contemplate the global-scale silver flow.

However, it is difficult to share, with non-Japanese scholarships, what is written in the old form of Japanese sources and the sources themselves, which makes people feel obscure to compare and think interconnections.

In order to think about the issues, thus, this project is not only to observe how to TEI-nize sources of multi languages, but also to feed-back how to regard problems occurring along with that such as the way to mark-up the units and technologies.

It is noteworthy that this project has a clear connection between the next paper by Naoki Kokaze, in terms of encoding transactions extant in historical sources. In that sense, this presentation deals with marking up basic information on mining in detail, based on the geographical and chronological descriptions.

Engi-shiki: Towards an Exemplary Markup Project in Japan

Presenter: **Naoki Kokaze**

The paper reports a joint research achievement through examples of TEI markup, based on 'Transactionography,' for an ancient Japanese historical source called Engi-shiki, which was compiled as an administrative manual around the tenth century C.E.. Ours is a cross-institutional research group, between the National Museum of Japanese History and University of Tokyo, that includes historians on ancient Japan and technical supporters.

One of our aims is to formulate a text database enabling historians to explore numerous questions by making use of various machine-processable data on society and administration in ancient Japan, according to their own interests. One similar research project, the Japanese Historical Text Initiative, based at the center for Japanese Studies at UC Berkeley, offers the Engi-shiki text written in Japanese and its English translation with the images of one of the manuscript volumes. Whereas that JHTI deals with the first 10 volumes of the Engi-shiki, our project will provide structured data of the full fifty volumes.

In this presentation, I will provide you with both of the brief introduction and prospective view of this project, mentioning the possibility of connecting TEI and IIIF.

[Day 5 (Thu) 10:45-12:15] LP Session VI (Room A1)
People, Persons, and Prosopography

TEI, the Walt Whitman Archive, and the Test of Time

Brett Barney[1]

The release notes for P5, version 2.0.0 TEI Guidelines (Nov. 2011) introduced new elements intended to "facilitate a more 'document-focussed' (as opposed to 'text-focussed') way of working" and a "mechanism . . . to represent the evolution of a text through various writing stages." My experiences in a project funded by National Endowment for the Humanities and Deutsche Forschungsgemeinschaft to test these new provisions suggest that considerable work remains before the TEI can claim to serve the needs of editors wishing to encode temporality. The Guidelines continue to rely on a flawed distinction between textual and non-textual features and provide insufficient guidance to those wishing to encode either fine-grained or inter-document diachronies.

Moreover, while <listChange> and its associated elements and attributes afford reasonably well-articulated methods for encoding "revision campaigns," my experiences in using them to encode Whitman manuscripts shows them to be inadequate for either the task of describing an editor's theories about the step-by-step inscription of a single document or for the task of encoding posited genetic relationships among documents. Ideally, provisions should allow encoders to specify discrete parts of documents as well as whole documents, to indicate gaps in the document record, and to hypothesize multiple, even mutually exclusive, possible genetic relationships.

The potential scholarly benefits of refining the markup of intra-document sequences are evident in several protypical interfaces created for the grant project. For inter-document diachronies, two possible basic frameworks are worth considering: RDF and "directed acyclic graphs," as originally suggested by the Workgroup on Genetic Editions.

References

Gerrit Brüning, Katrin Henzel, and Dietmar Pravida (2012): On the dual nature of written texts and its implications for the encoding of genetic manuscripts. In dh2012 – Book of Abstracts, Hamburg. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/on-the-dual-nature-of-written-texts-and-its-implications-for-the-encoding-of-genetic-manuscripts

Elena Pierazzo (2011): A rationale of digital documentary editions. In *Literary and Linguistic Computing* 26/4. 463–477. https://doi.org/10.1093/llc/fqr033

World Wide Women: TEI Prosopography and Global Genealogy in *Digital Dinah Craik*

Karen Bourrier[2], **Kailey Fukushima**[3]

In this paper, we explore the material conditions of digital editorial work that make uncovering nineteenth-century women's lives possible in the twenty-first century. Taking our project, *Digital Dinah Craik*, a TEI edition of the letters of the bestselling Victorian author,

[1] University of Nebraska-Lincoln, USA
[2] University of Calgary, Canada
[3] University of Victoria, Canada

as a case study, we discuss how TEI prosopography templates push us to extend our research practices beyond disciplinary and institutional borders. Following the work of Alison Booth and others, who demonstrate that prosopography can mitigate the erasure of women's lives, we combine digital scholarly editing with global genealogy in order to highlight the role of international women in our prosopography. We argue that combining research tools aimed at scholars, such as the TEI, with research tools aimed at general audiences, such as Ancestry.com and the British Newspaper Archive, leads us to a more global picture of women's and working-class histories. Our methods redress the academy's long tradition of suspicion towards what N. Katherine Hayles calls our "shadow fields" (36). Hayles writes that scholars often dismiss research activity conducted from outside the ivory tower "because it is not concerned with the questions the scholar regards as important or significant" (36). Although projects like the British Newspaper Archive are primarily aimed at family historians, as Patrick Leary points out, they are already transforming Victorianists' digital research practices. The mass digitization of newspaper and census material represents a "tidal deposit" of previously undiscoverable information "about the quotidian lives and beliefs of people in the nineteenth century" (Leary 268). We find that TEI prosopography templates push the boundaries of our research, spurring us to gather data to fill the tagsets for each person mentioned in our corpus, and in the process decentering power dynamics as unknown women travellers like the Brazilian-born Annie Miers get the same treatment as global literary celebrities.

Prosopography and Typological Analysis: Data Mining *Allgemeine Deutsche Biographie*

Tao Wang[1]

Allgemeine Deutsche Biographie (ADB) is an important tool for information about historical figures in the German-speaking world. The edition work was under the support of the Historische Kommission bei der Bayerischen Akademie der Wissenschaften and has been lasted for more than thirty years (1875-1912) for a total of 56 volumes with about 26,500 personal information collected. For a long time, the academic circles viewed the ADB only as reference book, which far underestimated the value of ADB. In fact, we can do some data mining in the context of Digital Humanities. The dataset we use here in our project benefits a lot from the technologies of TEI. We are not going to show how to deploy TEI, but to present how structured historical material with the help of TEI can be used for Information extraction.

Firstly, ADB can interpret historical figures with the idea of "Prosopography". Group biographies are particularly concerned with the trend of life expectancy. The entire ADB contains 26,000 historical figures, with an exact date of birth and death reaching 20,000. We can immediately calculate the average life expectancy, according to different time and to different occupations. We can see the different lifespan between historical celebrities and normal people. We can also see occupation is an important factor to affect life expectancy. To give these findings a reasonable explanation is challenging.

Another set of basic information about historical figures in the ADB is about where they were born and where they died. Based on the ADB's information, the movement of different historical figures from birth to death constitutes a fine network; with the aid of visualization tools, we can expose hidden information and outline a picture of the migration of historical figures, which deduced the formation of the center city of Germany. There is no doubt that there may be a relative concentration of birth and death, with the birth of more historical figures in certain places or the death of more people in certain locations. We present these factors as a weighted item in the visualization of the illustrations, produced a "death map" of the Germans and obtained interesting discoveries.

[1] Nanjing University, China

ADB's editorial team undoubtedly tried to emphasize that these locations are inextricably linked to Germany in history. The death map of the Germans also make us more intuitively aware that the competition between "*Großdeutsche Lösung*" and "*Kleindeutsche Lösung*" has a complicated historical fact. It is worth emphasizing that despite the existence of a relatively concentrated cities, the distribution and development of cities are relatively balanced across the entire German Empire.

Thirdly, as far as the overall occupation is concerned, ADB provided us with rich data from the 10th century onwards. Robert Merton, the American sociologist, used quantitative analysis to study the occupational status in England during 17th century. Merton tried to find the reason for the sudden emergence of British scientists in the 17th century. Through his macroscopic study on career change, he found that there was a connection between the accelerated development of British science and the transfer of professional interests of the elite in England at that time. We want to know, in the German context, the secular occupation changes can reflect the changing characteristics of the times? Is there any connection between the acceleration development of German society and the transfer of professional interests of the elite in Germany?

Figures are always the protagonists in promoting historical development. Biographical data will become an important clue for our understanding of history. Therefore, ADB has a special exemplary role, the academic community is also increasing emphasis on the processing of biographical dataset. When ADB project finished in 1912, a new project, an upgraded version of ADB, *Neue Deutsche Biographie* (NDB) launched in 1953. Subsequently, in the field of history there has also been attempts of biographical data to encompass a wider range of people, including the use of "relational" database thinking and the open-access idea for free sharing of information. With the support of the Deutsche *Forschungsgemeinschaft* (DFG), the German academic community launched a larger program called the "Deutsche Biographie" (DB), which not only expanded the "German" extension (in 2015, DB covered a total of 500,000 historical figures), but also on the technical form of the presentation biographical information, increasingly become the German world's important "Historical Information System." The biographical database represented by ADB is a gold mine for researchers and there are many hidden information worth exploring. Although our work has just begun, the present findings have opened another face for us in German history.

[Day 5 (Thu) 10:45-12:15] LP Session VII (Room A2)
Editing and Analysis

Genetic Encoding: A Reassessment with Lessons from the Faust Edition

Gerrit Brüning[1]

Around the same time as this year's TEI conference, version 1.0 of the Faust edition will be made open for the public. When the Faust edition (http://beta.faustedition.net/) was started in 2009, there were few approaches to genetic encoding. The project played an active role in the development of the "Encoding Model for Genetic Editions" (2010), most parts of which were integrated in chapter 11 of the TEI Guidelines in 2011.

With this first TEI P5 version 2 scholars face a variety of ways to represent witnesses. Besides new elements and attributes in chapter 11, version 2 allows for various encoding approaches that depend on how "document-focused" or how "text-focused" an encoder wishes an encoding to be. He or she may stick to the well-established structural elements and may or may not add some 'genetic' elements and attributes. Alternatively, an encoder might abandon the structural hierarchy and might instead embrace the spatially defined <sourceDoc>. The hierarchy of the traditional elements has no syntactical counterpart in <sourceDoc>, since its descendants <line> and <zone> may be nested freely (while in other respects the content model of the element <line> is strict).

As a result, even someone who is up-to-date with the current version of chapter 11 is unable to foresee how other scholars would encode a given document. It is, therefore, all the more import that we constantly discuss and compare our various approaches, current projects' experiences, challenges and solutions. This will be especially helpful for projects that are still in preparatory stage and want to make an educated decision between the various approaches. The talk pays attention to other notable projects such as the Shelley-Godwin Archive and addresses questions like: Do 'genetic' and "document-focused" encoding imply each other? What is "document-focused" encoding good for, what are its inherent weaknesses?

References

Brüning, Gerrit, Katrin Henzel, and Dietmar Pravida (2013) 'Multiple Encoding in Genetic Editions: The Case of "Faust"', *Journal of the Text Encoding Initiative,* 4, https://journals.openedition.org/jtei/697.

Muñoz, Trevor, and Raffaele Viglianti (2014) 'Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive', *Journal of the Text Encoding Initiative*, 8, https://journals.openedition.org/jtei/1270.

[1] Goethe University of Frankfurt, Germany

Métopes + TXM: Integrating Text Publishing and Text Analysis Tools Based on TEI Encoding

Alexei Lavrentiev[1], Charles Bourdot[2], Serge Heiden[3]

This paper presents an experience of creating workflows in text publishing and text corpus analysis projects that integrate, thanks to TEI encoding, two sets of tools created for different purposes.

The first set of tools is called Métopes (*Métodes et outils pour l'édition structurée*, or Methods and tools for structured publishing)[4]. It was developed by the *Pôle document numérique* of the Research centre for the Humanities (MRSH) in Caen (France) and consists in a full single-source publishing toolchain. After primary editing with Microsoft Word the documents are converted using special macros to TEI, which is the core format for further editing and for all publication forms, including PDF for printing (finalized with InDesign), ePubs and online editions produced dynamically from TEI sources by the MaX tool (based on BaseX)[5]. Métopes has been adopted by a number of French academic publishers.

TXM[6], on the other hand, is a free and open-source (GPL V3. licence) Java and C based platform for text corpus building, annotation and analysis. It includes NLP tools, search engines and visualization tools with convenient hyperlinks between distant synthetic quantitative analysis to close reading views . TXM uses TEI (with a couple of extension elements)[7] as an internal format for encoding the text structure and all kinds of annotations.

So, both Métopes and TXM rely on TEI markup. However, Métopes focuses on general text structure and on presentational aspects (e.g. it is very sensitive to white spaces), while TXM needs to perform precise linguistic analysis (e.g. tokenisation, language identification in multi-language documents).

Thanks to funding from CAHIER consortium[8], an intern from Caen worked for three months with the TXM team in 2017. He created a set of XSLT and CSS stylesheets that make it possible to correctly parse and analyse a Métopes produced text file or corpus with TXM, and to generate high quality publications based on texts prepared for TXM analysis. In many cases both tools use the same TEI tags, which makes integration quite straightforward. In other cases, more work is necessary to ensure full compatibility (e.g. generating table of contents in TXM or supporting word-level annotation by Métopes tools). A further integration step may consist in creating a single editorial and analytical toolchain for text scholars. A simplified workflow chart of this chain is presented in the *Figure 1*.

The work done so far is documented on a wiki page[9], and the scripts (Groovy and XSLT) are available for download under an LGPL license. All documentation is currently only available in French but we are interested in collaboration for its translation into English and other languages.

In the presentation at the conference we will provide on methodological aspects of combining analytic and editorial markup, and will show examples of works prepared using both TXM and Métopes.

---

[1] CNRS, France
[2] Université d'Angers, France
[3] ENS de Lyon, Frace
[4] http://www.unicaen.fr/recherche/mrsh/document_numerique/outils/metopes
[5] http://www.unicaen.fr/recherche/mrsh/document_numerique/outils/max
[6] http://textometrie.org
[7] https://wiki.tei-c.org/index.php/TXM
[8] https://groupes.renater.fr/wiki/txm-info/public/xml_tei_txm
[9] https://groupes.renater.fr/wiki/txm-users/public/usr_mrsh

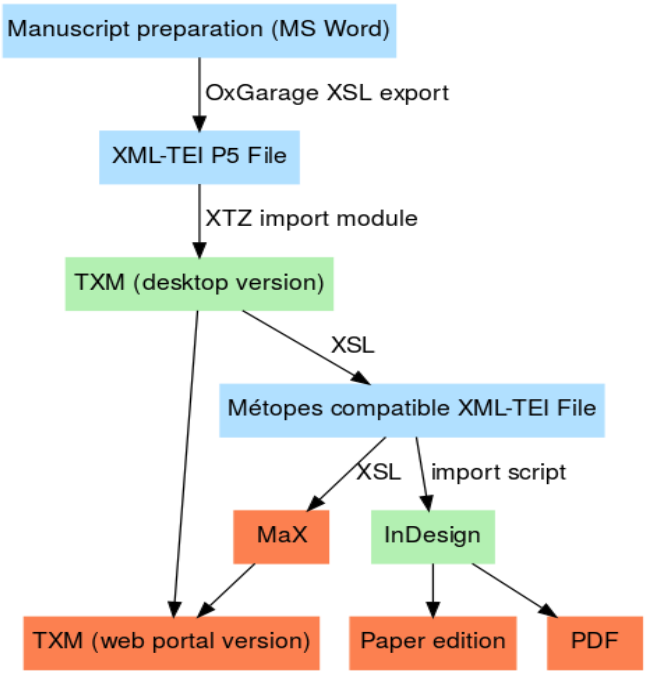


Figure 1. Text analysis and publication workflow integrating Métopes and TXM

Quotes, Paraphrases, and Allusions: Text-reuse in Sanskrit Commentaries and How to Encode it

Patrick McAllister[1]

The Sanskrit literary genre of commentaries has several characteristics that are a challenge to the structural encoding of its texts. One particular tricky feature is the skillful and sophisticated reuse of texts in such commentaries: whilst one can easily distinguish different types of (what we currently call) 'quotes'—literal ones, paraphrases, allusions, summaries—it is often helpful also to differentiate the voice in which the text is reused (that of an opponent, an ally, or someone in between), as well as to note the varying degrees of authority and acceptance that are attached to a 'quote'.

The talk will present several practical examples of encoding these types of passages, drawing on the documents of SARIT ("Search and Retrieval of Indic Texts", http://sarit.indology.info) and the presenter's own encoding projects. The discussion of the advantages and disadvantages of the various approaches will consider three related but distinct purposes that this encoding should serve: to allow an editor to judge textual variation between quoted source and its quote, to allow a reader to glean helpful hints for understanding the text, and to allow for an easy extraction and presentation of the encoded information.

[1] Institute for the Cultural and Intellectual History of Asia (IKGA), Austrian Academy of Sciences, Austria

[Day 5 (Thu) 10:45-12:15] SP Session III (Room B1,2,3)

On Global, Formal Languages, and the Others

Susanna Allés-Torrent[1], Mitsunori Ogihara[1]

Our contribution arises from our experience on creating undergraduate DH courses, and the challenges encountered to balance the use of TEI, as a Global language, and its integration in different pedagogical contexts, namely in the foreign language classroom. While sometimes the nature of the languages studied differs —formal and artificial languages encapsulate human thoughts and information processing; natural languages are the instinctive expressions that materialize human thinking—, in both scenarios we deal with the ultimate goal of using a communication system to establish interactions (human-human, human-machine, and machine-machine). In late debates on Global-Local, North-South, or Center-Periphery, the TEI serves as a paradigmatic case, allowing for many years the interchange of texts, with a well-deserved international impact. When teaching TEI students learn a different approach to texts, they consider the textual structure and the semantics of the text and they translate it in a machine-readable form (Battershill & Ross 2017; Rehbein & Fritze 2012). Learning TEI gives the students the skills for annotating any kind of document with a shared syntax and with specific semantics, and for sharing any type of data information with other users while minimizing misunderstandings. TEI's global dimension, however, can sometimes be challenged by practice and specific contexts (Dee 2014). We will offer different scenarios and solutions, including teaching TEI in foreign language classrooms (Spanish), where the students must learn, at the same time, a foreign natural language and a global language for its annotation. We present some strategies where TEI empowers the interpretative processes for cultural and literary texts, conceiving TEI as a way of close reading for the appropriation both of the literary meaning and of a foreign second language. We also report some initiatives undertaken to solve the dearth of non-English resources, such as the creation of online *ad hoc* pedagogical materials, and an initiative to revise and complete the Spanish translation of the Guidelines.

References

Battershill, Claire, and Shawna Ross (2017) *Using Digital Humanities in the Classroom*. A Practical Introduction for Teachers, Lecturers, and Students. London; New York: Bloomsbury.

Dee, Stella (2014) "Learning the TEI in a Digital Environment." *Journal of the Text Encoding Initiative* 7 (November). doi.org/10.4000/jtei.968

Rehbein, Malte, and Christiane Fritze (2012) "Hands-On Teaching Digital Humanities: A Didactic Analysis of a Summer School Course on Digital Editing". *Digital Humanities Pedagogy. Practices, Principles and Politics*, ed. by Brett Hirsch. Cambridge: Open Book Publishers. doi.org/10.11647/OBP.0024.

On Structuralism and the Predictive Attribution of Type

Vinayak Das Gupta[2]

It is well established that the Text Encoding Initiative (TEI) follows a model that assumes that a text is hierarchical – that it is an ordered hierarchy of content objects (Renear 1993; Hayles 2005; Cummings 2008). In other words, TEI resists the poststructuralist turn,

[1] University of Miami, USA
[2] Shiv Nadar University, India

proposing an older, structuralist view of the written word. This proposed scientific objectivity in the realm of literary studies, by subordinating 'parole' to 'langue' (Bakhtin), is thought to be of significance in creating a machine-readable (and human-readable) document. Such a structuralist perspective foregrounds the patterns, systems, and structures of a text over the specific content. If we are to agree that a particular *form* of a text (for instance, a poem, a novel, or a play) has specific components (the basis on which TEI schemata exist) that are contingent on the form, we can argue that if certain components of a text appear in a certain order, it is of a *particular* form. This follows from the writings of Levi-Strauss proposing that structures are universal, hence ahistorical[1].

The *Letters of 1916* are born from crowd-sourced transcriptions which are given some markers (in the form of tags); could the structure of the text -- as made distinct by specific markers like <address>, <date>, <salutation>, <signed>, <pb>, <p> etc. – be used to predict the form of the document (letter, postcard etc.)? This paper produces a theoretical argument connecting structuralism and how it plays a key role in our attempts to identify the form of a document.

References

Bakhtin, M. (1981). *The Dialogic Imagination.* Austin: University of Texas Press.

Cummings, J. (2008). 'The Text Encoding Initiative and the Study of Literature.' *A Companion to Digital Literary Studies*, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell.

Hayles, N. Katherine (2005). *My Mother Was a Computer: Digital Subjects and Literary Texts*. Chicago: University of Chicago Press.

McGann, J. (2001). *Radiant Textuality: Literature After the World Wide Web*. Palgrave.

Renear A., E. Mylonas, and D. Durand (1993). *Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies*. <http://www.stg.brown.edu/resources/stg/monographs/ohco.html>. Accessed April 19, 2018. <http://www.digitalhumanities.org/companionDLS/> Accessed April 19, 2018.

Encoding GeoJSON Geometries in TEI

Martin Holmes[2]

Abstract

This presentation will address two distinct but related issues concerning the representation of feature geometries on two-dimensional surfaces in TEI.

Over the last decade, GeoJSON has proved itself as a simple, efficient and rich standard for encoding map features, and is widely used throughout the GIS community, eclipsing other apparently more XML-friendly options such as GML and KML. As more TEI projects incorporate GIS data, there is a need to provide examples of good practice which enable the integration of GeoJSON-encoded feature geometries (Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon and MultiGeometry) with TEI encoding. The first half of this presentation will show how the University of Victoria's BreezeMap project is approaching this problem.

The second half of the presentation will address the need to encode similar complex geometries inside the TEI <facsimile> element, to represent points, lines and

[1] This paper is aware of the debate surrounding the issues with hierarchical organisation of text as proposed by Jacques Derrida, or, in the case of TEI, by Jerome McGann. McGann contends that 'its [XML's] hierarchical principles and other design characteristics set permanent and unacceptable limits on its usefulness with arts and humanities materials.' (17: 2001). This paper will attempt to engage with this area of inquiry to some extent.

[2] University of Victoria, Canada

polygons on text-bearing surfaces or other two-dimensional coordinate spaces. The model of GeoJSON can be helpful here too, and although the existing TEI element-set (<surface>, <zone>, and the soon-to-be-added <path>), along with their current attribute array, already make this possible, there are many possible approaches that might be taken for any given geometry. It would be desirable to settle on a set of best practices and exemplify them in the TEI Guidelines.

1. GeoJSON and BreezeMap

Although the GeoJSON format was both stable and widely supported by 2008, it was not until 2016 that the Internet Engineering Task Force adopted it, with some minor changes, as an RFC (IETF 2016). Software support for GeoJSON is now virtually universal in the mapping community.

GeoJSON provides a clear and easily-processable array of feature geometries, including primitives (Point, LineString, Polygon) and complex combinations (MultiPoint, MultiLineString, MultiPolygon and MultiGeometry) to enable the encoding of locations (features) on a mapping surface using GIS coordinates.

Our BreezeMap project aims to create a toolset based on OpenLayers, GeoJSON and TEI for creating interactive maps and gazetteers, as well as an approach for encoding complex features on textual facsimiles and other surfaces represented as two-dimensional digital images. This presentation will discuss the approaches taken to integrating TEI and GeoJSON in the BreezeMap project.

2. Mapping in TEI

Projects marrying TEI and GIS are increasingly common (Hickcox et al 2013, Jenstad 2006-present). The TEI treats places as first-order entities, and building a gazetteer using <listPlace> and <place> is straightforward, but support for actual GIS is currently very limited (see my TEI feature request Need to improve GIS/gazetteer encoding support in TEI). The TEI <geo> element "contains any expression of a set of geographic coordinates, representing a point, line, or area on the surface of the earth in some notation." (TEI 2018), a rather vague definition which is currently supplemented by only a few examples showing single points encoded as a simple coordinate pair, or embedded GML showing a Linear Ring. For encoders who would like to make use of the range of geometries available in modern mapping standards, it would help to have a few clear examples showing GeoJSON embedded in TEI, as in Example 1, which is taken from a campus map that forms one of the test projects in BreezeMap:

```
<place xml:id="bldgElliottTheatre" corresp="#bldgTeaching">
 <placeName>Elliott Lecture Theatre</placeName>
 <desc>...</desc>
 <location type="GeoJSON">
  <geo crs="urn:ogc:def:crs:OGC:1.3:CRS84"
geoEncoding="GeoJSON">"geometry":{"type":"Polygon",
   "coordinates":[[[-123.3108016,48.4627526],[-123.3102577,48.4627937],[-
123.3102336,48.4626537],[-123.3102028,48.4624747],[-123.3107466,48.4624335],[-
123.310778,48.4626155],[-123.3108016,48.4627526]]]}</geo>
 </location>
</place>
```

Example 1. A TEI <place> element incorporating GeoJSON geometry.

The @crs and @geoEncoding attributes, which are part of the feature request mentioned above, will provide better options for machine-readable encoding than the current <geoDecl> header element allows. The actual feature geometry is expressed in its pure GeoJSON format, and is easily machine-readable. Converting a TEI gazetteer constructed like this into a GeoJSON file for display in a mapping application is trivial, especially with XSLT 3.0's enhanced support for JSON.

3. Facsimiles

The GeoJSON RFC specifies that "GeoJSON uses a geographic coordinate reference system, World Geodetic System 1984, and units of decimal degrees.," and "the first two elements [of a position] are longitude and latitude, or easting and northing," (IETF 2016); in other words, it is intended only for geographic data. However, there is nothing intrinsic to the specification of geometries in the RFC which precludes using the same structures to describe features on any two- or three-dimensional surface. The same types of feature (points, lines, polygons, and combinations of them) are a natural fit when annotating manuscript images, engravings, and other pictorial or textual content in any TEI project which incorporates a facsimile component. A mapping library such as OpenLayers will happily display any static image with features described in "GeoJSON" which actually consist only of pixel offsets. This is the approach we take when encoding pseudo-geographical coordinates on historical maps which are not, or cannot be, geo-rectified, as in the case of the Agas Map (Jenstad et al. 2015).

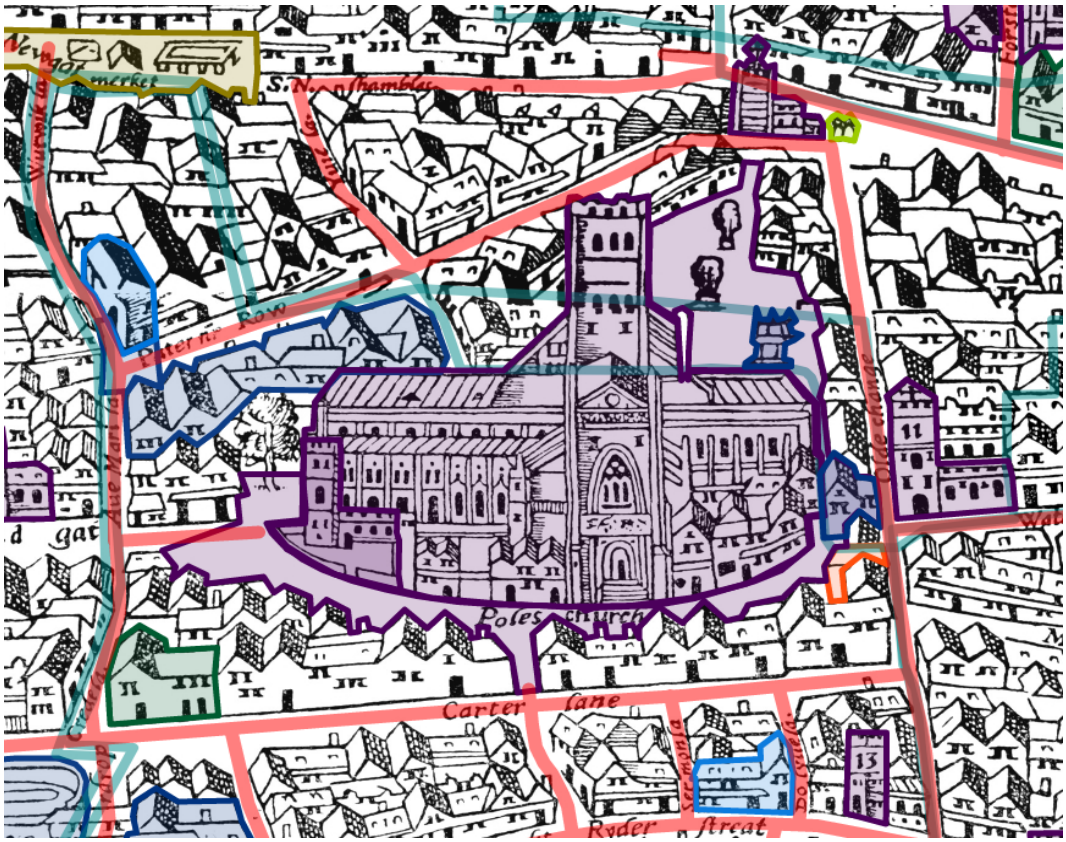


Figure 1. A detail from the Agas Map, which shows features outlines using polygons and lines which are expressed in GeoJSON in the TEI encoding.

This makes "GeoJSON" (or at least, its formal structure) a natural fit for rendering annotated images, and interfaces such as those provided by OpenLayers drawing tools a good option for "authoring" such features. The question then is how best to encode the resulting data, in a non-lossy manner, in the existing TEI elements for describing features on a surface (<zone> and the soon-to-be-available <path>). Browser-based drawing tools developed and optimized for maps can now provide a more powerful alternative to older, less sophisticated tools for facsimile editing, such as my own Image Markup Tool (Holmes 2012).

The remainder of this presentation will examine ways in which such geometries can be encoded in TEI, and easily recovered into GeoJSON structures for rendering purposes.

Keywords
maps, GeoJSON, facsimiles

References

Hickcox, Alice, Michael C. Page, and Randy Gue (2013) "The Atlanta Map project: TEI and GIS collaborate to create a research environment." Digital Humanities 2013 Conference, Nebraska. http://dh2013.unl.edu/abstracts/ab-416.html.

Holmes, Martin (2012) *Image Markup Tool*. Version 1.8.2.2. https://hcmc.uvic.ca/~mholmes/image_markup/.

Internet Engineering Task Force (2016) *The GeoJSON Format*. Request for Comments (RFC) No. 7946. https://tools.ietf.org/html/rfc7946.

Jenstad, Janelle, Kim McLean-Fiander, Greg Newton and Martin Holmes (2015) "How To Edit a Map in TEI." Digital Humanities 2015 Conference, Sydney, Australia. http://dh2015.org/abstracts/xml/JENSTAD_Janelle_Auriol_How_To_Edit_a_Map_in_TEI//JENSTAD_Janelle_Auriol_How_To_Edit_a_Map_in_TEI.html.

Jenstad, Janelle. 2006-present. The Map of Early Modern London. http://mapoflondon.uvic.ca/

TEI Consortium (2018) TEI Guidelines, version 3.3.0.

Automatically Encoding Encyclopedic-like Resources in TEI

Mohamed Khemakhem[1,2,3], Laurent Romary[1,2,4], Simon Gabay[5], Hervé Bohbot[6], Francesca Frontini[6], Giancarlo Luxardo[6]

Encyclopedic content exists in different forms of paper-based resources and remains to a very large extent not exploited given the limited alternatives to extract information from the corresponding digitized material. Where TEI has provided mechanisms for encoding such a content, recent works have yielded advanced techniques to apply these mechanisms automatically on raw resources, however clear recommendations for the encoding of such content is still lacking.

As current digitization projects concern encyclopedias, other projects are undertaken to digitize textual resources with heritage significance and with similar features in terms of layout and indexed access. In this paper we introduce new lexicographic models, customizing the TEI schema, we extend existing ones and we present their application on two categories of resources, currently under development to be made available for public use:

- Encyclopedic sections of early, out of copyright, versions of a reference French Dictionary (Petit Larousse Illustré),
- Manuscripts auction catalogues, which are old resources for referencing and semantically describing manuscripts for sale.

The definition of these models is based on the significant similarities noticed at different levels of the described information in such entry-based resources. We have employed elements from the TEI dictionaries module to enable the segmentation of morphological and semantic information encompassed by respectively <form> and <sense> elements, already extracted by dedicated existing models in GROBID-Dictionaries[7]'s architecture. In fact, each entry in both types of resources carries two main encyclopedic information. First,

[1] Inria ALMAnaCH, France
[2] Centre Marc Bloch, Germany
[3] Paris Diderot University, France
[4] Berlin-Brandenburgische Akademie der Wissenschaften, Germany
[5] Université de Neuchâtel - Institut de littérature française, France
[6] Univ Paul Valéry Montpellier 3, CNRS, PRAXILING UMR 5267, France
[7] https://github.com/MedKhem/grobid-dictionaries/

the name or the label of a concept, coupled with one or more of its extended form and coming sometimes with a brief description. Then, a second block comes to establish an exhaustive description of the concept and to present related information such as bibliography or pricing. In addition, we have used a customized version of the <entry> element to allow the generic encoding of entry numbering (see second excerpt in the table below).

TEI encoding	Excerpt from Dictionary Encyclopedic Section in Petit Larousse Illustré (1948)
<entry> <form type="lemma"> <persName>ABERDEEN</persName> <addName>(G. H. Gordon, comte d')</addName><pc>,</pc> <desc>homme d'Etat anglais, né à Edimbourg</desc> </form> <pc>.</pc> <sense> <def>Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860)</def> </sense> <pc>.</pc> </entry>	**ABERDEEN** [*aberdin'*], v. d'Ecosse, ch.-l. de comté; port sur la mer du Nord; 170.000 h. Université. **ABERDEEN** (G. H. Gordon, *comte d'*), homme d'Etat anglais, né à Edimbourg. Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860). **ABER-VRACH**, fl. côtier du Finistère (Atlantique); 34 kil. Station marémotrice d'essai. **ABGAR**, nom de huit rois d'Edesse, en Mésopotamie (132 av. J.-C.-216 apr.). **ABIA**, roi de Juda, fils de Roboam, vainqueur de Jéroboam, roi d'Israël (957-955 av. J.-C.). **ABIDJAN**, ch.-l. de la Côte-d'Ivoire (A.-O. F.), sur une vaste lagune navigable; 15.000 h. **ABIMÉLECH** [*lèk*], fils de Gédéon. Il devint Juge d'Israël, après avoir fait égorger ses frères; il établit son pouvoir sur Sichem et fut tué au siège de Thèbes, en Palestine (vers 1100 av. J.-C.). **ABIRON**, lévite qui fut englouti dans la terre avec Coré et Dathan, tous trois révoltés contre Moïse et Aaron (*Bible*).

TEI encoding	Excerpt from a Manuscripts Auction Catalogue (1889)
<entry> <num>49</num> <form type="lemma"> <surName>Kourakin</surName> <addName>(le prince Alexis B.),</addName> <desc> frère du précédent, homme d'Etat russe.</desc> </form> <sense> <pc>-</pc> <def> <bibl>Billet auto sig., en francais, à M. Monférand, 1 p, in-8.</bibl> <num type="price">2 »</num> </def> </sense> </entry> <entry> <num>54</num> <form type="lemma"> <surname>Lassalle</surname> <addName>(A.-Ch.-L. de)</addName>, <desc>le plus brillant général de cavalerie des guerres de la République et de l'Emplre, né à Metz, tué à la bataille de Wagram</desc> </form> <sense> .- <def> <bibl>L. a. s. au général Dugua; 1 p. in-f.</bibl> <num type="price">10 »</num> </def> <note>Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite affaire, tué 10 mameloucks et 40 arabes, etc.</note> </sense> </entry>	49 **Kourakin** (le prince Alexis B.), frère du précédent, homme d'Etat russe. — Billet aut. sig., en français, à M. Monférand, 1 p. in-8. 2 » 50 **Labanoff** (le prince Alex.), célèbre général et écrivain russe, historien de Marie Stuart. — L. a. s., en français, 1835, 1 p. in-4. 3 » 51 **Ladislas IV,** roi de Pologne, célèbre par ses succès contre les Russes, époux de Marie de Gonzague. — L. sig., en latin, au cardinal de Montalte; Varsovie, 1645, 1 p. in-f. 8 » 52 **Lafayette,** illustre général. — L. a. sig. de ses initiales à M. Masclet; Washington, 13 août 1825, 1 p. 1/4 in-4. Un peu fatiguée. 15 » Très-curieuse lettre sur le voyage qu'il fit en Amérique, de 1824 à 1825. « C'est avec de bien tendres regrets que je quitterai cette terre américaine, le bon, grand et heureux peuple des Etats-Unis auquel je suis amalgamé depuis près d'un demi-siècle, et qui vient encore de me combler de ses bontés. J'y ai vu les miracles de l'indépendance, de la liberté, égalité et *self government;* le problème des institutions républicaines a été résolu ici sur une grande échelle et jamais expérience n'a si bien réussi. » Il comptait retourner comme il était venu, sur un paquebot-poste, mais le peuple et le gouvernement en ont disposé autrement. On a donné le nom de *Brandywine* à une superbe frégate qui est chargée de le ramener en France. 53 **La Roncière** (Emile-*Clément* de), fils du général, condamné pour tentative de viol. — L. a. s. aux officiers et élèves de l'école de Saumur; Paris, mai 1836, 3 p. pet. in-4. 10 » Très-curieuse lettre toute relative à son procès. 54 **Lassalle** (A.-Ch.-L. de), le plus brillant général de cavalerie des guerres de la République et de l'Empire, né à Metz, tué à la bataille de Wagram. — L. a. s. au général Dugua; 1 p. in-f. 10 » Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite affaire, tué 10 mameloucks et 40 arabes, etc.

We employed the adapted and newly defined models to automatically process and label the encyclopedic information in a cascading fashion, as introduced in Khemakhem et al. 2017 and Khemakhem et al. 2018, by relying on the same sequence labeling machine learning technique.

We will present in-depth our progress in implementing morphological and semantic models and show through the results at each structuring level how such information could be uniformly encoded and automatically extracted from these resources.

References

Mohamed Khemakhem, Luca Foppiano, Laurent Romary (2017) Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. electronic lexicography, eLex 2017, Sep, Leiden, Netherlands

Mohamed Khemakhem, Axel Herold, Laurent Romary (2018) Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018, May, Miyazaki, Japan.

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem, Laurent Romary (2018) Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. GLOBALEX 2018 - Globalex workshop at LREC 2018, May, Miyazaki, Japan.

[Day 5 (Thu) 13:15-14:45] LP Session VIII (Room A1)
Encoding: Etymology, Liturgy, and Festivals

Encoding Mixtepec-Mixtec Etymology in TEI

Jack Bowers[1], Laurent Romary[2]

In this paper we present the encoding of Mixtepec-Mixtec language (iso 639-3: mix) etymological information in TEI which is one component of a larger ongoing language documentation project (Bowers & Romary, 2017a). Mixtepec-Mixtec is an under-resourced Otomonguean language spoken in Juxtlahuaca district of Oaxaca, and parts of the Guerrero and Puebla states of Mexico.

Though no variety of Mixtecan has any orthographic or phonetic attestation before 1567[3,4], there is nonetheless a great deal we can see with regards to the origin of a wide variety of the vocabulary, particularly by means of: polysemy, compounds, inference via anthropological knowledge and some evidence is even visible in the various pictographic Mixtec codexes[5].

Mixtecan semantics, polysemy and lexical innovation have been the topic of a number of important studies (cf. Brugman,1983; Brugman and Macaulay,1986; Hollenbach, 1995). These works provide a number of examples (particularly from body-part terms and spatial terms) which bring to light certain patterns observable in language change which show clear evidence for underlying conceptual and cognitive motivations and mechanisms involved. The data from Mixtepec-Mixtec adds to this discussion. Specific phenomena observed and encoded in the dataset include: borrowing, inheritance (in conjunction with cognates from related varieties of Mixtec), derivation, compounding, various sense changes such as: metaphor and metonomy, and grammaticalization.

Thus, herein we will introduce some of the more interesting observations in Mixtepec-Mixtec etymology while demonstrating the application of recommendations put forth for the encoding of etymological information in TEI of Bowers & Romary (2017b), Sagot (2017), as well as expansions to the former as per Bowers et al (forthcoming).

References

Bowers J. and L. Romary (2017a) "Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec," *JADH 2017: Proceedings of the 7th Conference of Japanese Association for Digital Humanities "Creating Data through Collaboration"*, Sep, Kyoto, Japan. 〈hal-01744813〉

Bowers J. and L. Romary (2017b). "Deep encoding of etymological information in TEI," *Journal of the Text Encoding Initiative*, TEI Consortium, 〈https://jtei.revues.org/1643〉. 〈10.4000/jtei.1643〉. 〈hal-01296498v2〉

Bowers J., B. Sagot and L. Romary. Unpublished manuscript. "Going deeper into the encoding of etymology in TEI".

Brugman, C. (1983). "The use of body-part terms as locatives in Chalcatongo Mixtec," *Survey of California and Other Indian Language*s, 4, 239–290.

[1] Austrian Center for Digital Humanities (ACDH), Austrian Academy of Sciences (ÖAW), Austria

[2] Inria, France

[3] Hernández, Fray Benito. 1567. Doctrina Christiana en Lengua Mixteca. México: Casa de Pedro Ocharte.

[4] The varieties in attested in the above were from Tlaxiaco and Ayutla

[5] It should be noted though the content of the codexes are not language specific and are not unique to the Mixtepec variety.

Brugman, C., & Macaulay, M. (1986). "Interacting semantic systems: Mixtec expressions of location," In Annual Meeting of the Berkeley Linguistics Society (Vol. 12, pp. 315–327). Retrieved from http://journals.linguisticsociety.org/proceedings/index.php/BLS/article/download/3179/2898

Hollenbach, B. E. (1995). "Semantic and Syntactic Extensions of Body-Part Terms in Mextecan: The Case of "Face" and "Foot"," *International Journal of American Linguistics*, 168–190.

Sagot B. (2017). "Extracting an Etymological Database from Wiktionary," *Electronic Lexicography in the 21st century (eLex 2017)*, Sep, Leiden, Netherlands. pp.716-728, 〈https://elex.link/elex2017/〉. 〈hal-01592061〉

Encoding Liturgy - from the Haggadah to a General Schema

Yael Netzer[1], Sinai Rusinek[2], Andrew Irving[3], Clemens Leonhard[4]

Liturgical texts are an endeavor to standardize ritualized behavior in a context of a tradition which is normative, yet fluid. Such texts pose various challenges to their encoding in TEI, the most baffling of which are related to their performative nature. They prescribe a sequence of actions, gestures, and postures; some contain the utterance of words.
Scribes of liturgical texts avail themselves of intricate cues - textual and visual - to demarcate prescriptions of utterances and of actions. Yet, relying on implicit knowledge of the practitioners of liturgy, these cues are often inconsistent, and may change even within one document, for example, for aesthetic considerations.

Our primary dilemma is: should encoding keep faithful to the concrete manuscripts as an object in front of the readers, with its codicological and palaeographical features, or rather, our own understanding of the textual structure and the authorial intentions? TEI severely restricts the application of different grids and layers of structure within one text.

Our chosen text, the Haggadah, is read annually at the eve of passover in Jewish families since at least the Middle Ages. It stages the fulfilment of the duty of the head of the family or the presider of a group to transfer the memory of liberation from slavery in Egypt to the family's children, and to discuss the historical identity of the people with the participants. Like other liturgical texts such as the Christian Mass, its performance represents bits of collective memory of the ceremony.

This is not the first attempt to represent liturgical text within the TEI framework[5]. Conversing with our predecessors, we do propose a set of liturgical tags, such as <liturgUnit> <instructions> and <liturgFormula>, as part of a schema that will enable a better representation of liturgical texts of various types and of their unique features and idiosyncrasies.

---

[1] Ben Gurion University, Dicta, Tel Aviv University, Israel
[2] Haifa University, Bar Ilan University, Open University, Israel
[3] University of Groningen, Netherland
[4] University of Münster, Germany
[5] James Cummings: CURSUS An On-line Resource of Medieval Liturgical Texts. <http://www.arts-humanities.net/node/2113> created: 2007-07-20, last updated 2011-05-11 15:38. Idem, (2006). Liturgy, Drama, and the Archive: Three conversions from legacy formats to TEI XML. Digital Medievalist. 2. DOI: http://doi.org/10.16995/dm.11

Documentation and Digitisation of Festival in Pelu Awofeso's *White Lagos: A Definitive and Visual Guide to the Eyo Festival*

Felix Bayode Oke[1]

Festivals are significant events in the social and cultural reality of a people. To preserve cultural heritage, specialists capture what happens before, during, and after a festival by interviewing participants, taking photographs and record audio and video of the event, etc. For example, Pelu Awofeso has documented the Lagos Eyo Festival (also known as the Adamo Orisha Play) in his work *White Lagos: A Definitive and Visual Guide to the Eyo Festival,* in which he observes participants and uses a narratological approach to document the event in textual form. In this paper, I argue that the use of TEI, with its tagging for recording events and dialogue between speakers and its semantic markup for participants, places, and organisations, will enhance the preservation, data protection, and privacy of textual documentation of festivals.

Keywords
Digitisation, Eyo Festival, cultural heritage, Text Encoding Initiative, Digital Festival

[1] Anchor University Lagos, Nigeria

[Day 5 (Thu) 13:15-14:45] LP Session IX (Room A2)
Encoding: Newspapers, Commentaries, and Manuscripts

Creating a Digital Newspaper Collection in Dialogue with its Users

Dario Kampkaspar[1], Claudia Resch[1]

At last year's TEI conference, we presented the Wiennerisches Diarium, one of the oldest newspapers in the world. The project aims at providing (in a first step) 5 issues for almost every year between the first issue in 1703 and 1799. While the full text is provided by automatic text recognition and subsequent proofreading, annotations to the text – free text, labels for a paragraph, and identification of a person, place, event or other phenomena of interest – will be made by the users themselves.

In order to better understand our user's needs, we have involved interested scientists from different disciplines in two "annotate-a-thons" and, in the course of a two day conference in April, met those who already use the "Diarium" and those who are interested in doing so to discuss how a digital edition of historical newspapers should ideally look like to lend itself to research scenarios from a range of different disciplines. Participants at the conference were asked to fill in a questionnaire about functionalities needed for their research and what kinds of annotations they want to make. A final analysis, however, has yet to be done and will be presented at the conference.

An application is currently pending for a grant to deepen the research into how users interact with a digital edition and how the interface should be designed to make this interaction as easy as possible. The Diarium framework will implement these findings as far as possible. Its current features include administrative functions, full text search and user annotations. All user-supplied information is kept separate from the text. While still under development, a release of the framework is planned for summer. The project will share its findings, especially the results of the questionnaire and design research, via TEI-L and the SIG newspapers and periodicals.

How to Encode the Tibetan Commentaries on the *Abhidharmasamuccaya*

Koichi Takahashi[2], Hiroshi Nemoto[3]

This presentation introduces our project researching on the classical Tibetan commentaries on the *Abhidharmasamuccaya,* a glossary of the Buddhist technical terms. It was written in Sanskrit about 5C, and translated into Tibetan about 8C. This work arguably didn't draw attention in India, but Tibetan Buddhist scholars seemed to regard it as an important work. Especially Tibetan monks belonging to a group called "bKa' gdams pa" had written various commentaries on it since 11C. The works of bKa' gdams group were found at the beginning of 2000s, and the replicas of them have been published as the collected works of *bKa' gdams pa* or *bKa' gdam gsum 'bum* since 2006. In this collection, more than 10 commentaries on the *Abhidharmasamuccaya* are included. Our project focuses on these new materials and aims to encode them by using TEI P5 to develop a machine readable text for the field of Tibetan Buddhist studies. The format of

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria
[2] The University of Tokyo, Japan
[3] Hiroshima University, Japan

text in the collected works of bKa' gdams pa has a traditional style of the Tibetan classical literature. The book in the traditional Tibetan style looks like a bundle of rectangular paper. In the case of commentaries of the *Abhidharmasamuccaya,* each one consists of about 100 or 200 sheets of paper. Both sides of a sheet have 6 or 8 lines written from left to right without showing the head of a paragraph by breaking lines. Firstly, our project will encode the physical structures of the text by <pb /> and <lb />. Then, we attempts to mark up the logical structures like a paragraph by <p>. Added to that, we try to show the relationship between the *Abhidharmasamuccaya*, the commented work, and the commentaries by means of <xr>.

Beta maṣāḥəft: Encoding Ethiopic Manuscripts in TEI

Pietro Maria Liuzzo[1]

Beta maṣāḥəft, Manuscripts of Ethiopia and Eritrea (Beta maṣāḥəft means library) is a research environment. But what does 'research environment' mean? Our take is that a research environment for the cataloguing of Ethiopian manuscripts, encoding of digital editions of literary works and many other resources about Ethiopia should not just be developed as a data entry, management and editing tool packed with specific features, but as an entirely flexible and expandable system based on solid standards, like TEI and RDF. It should expand in terms of interests covered, in terms of participants and still stay focused enough in its presentations as to allow specific fields of research. Participants should not have to enter what we encode, they should be able to contribute what they want to encode and meet enough design and architecture flexibility to join easily, quickly and well. Also the web views and other outputs should not be centrally decided but in continuous development according to the needs of participating members, and there should be enough access points to the data in as many different ways as to allow anyone to build applications on that data. We think a TEI based workflow where all participants have enough knowledge of the project schema serves this perfectly and allows the extensibility of the encoding in the direction required from time to time providing at each collaboration benefits for all. Additionally, we use TEI as basis for IIIF manifests and to produce RDF for several purposes. In this paper we would like to bring our concrete example of a world scale TEI project about Ethiopian and Eritrean Manuscript culture putting more effort on collaboration about the data encoding then on development of views and websites, allowing extra schema flexibility and enforcing strict documentation policies.

[1] Universität Hamburg, Germany

[Day 5 (Thu) 13:15-14:45] Panel Session III (Room B1,2,3)

Implementing TEI to Japanese Modern Texts

Yu Okubo[1], Yoko Mabuchi[2], Kiyonori Nagasaki[3]

In Japan, TEI has recently been adopted in several projects to encode Japanese modern texts. Due to several reasons, in the past, and still now, other markup schemes had been in usage. Recently, some technical issues have been solved such as compatibilities between languages. However, attempts with TEI reveals intrinsic difficulties in the encoding of Japanese texts. It is especially difficult to encode in TEI pre-modern texts distributed in woodcut printings and manuscripts, despite the fact that it is relatively easy to mark up texts printed by metal type and the born digital.

Given this situation, this panel offers three presentations that seek for possibilities to adopt TEI to Japanese modern texts. First, Prof. Yu Okubo will present the current situation and workflow of Aozora-Bunko to transcribe, proofread, and encode out-of-copyright Japanese texts according to their original markup schemes, so that we will share the collaborative markup for Japanese through its experience. Second, Dr. Yoko Mabuchi will present an account of efforts to mark up Japanese historical corpora as well. Dr. Kiyonori Nagasaki will make a presentation of a use case of TEI on a collaborative parallel text database. The panel will be followed by general discussion.

Aozora Chuki: A Japanese Markup Language for Digitalized Texts in Aozora Bunko
Presenter: **Yu Okubo**

Aozora Bunko (Open Air Library), a non-profit digital archive for Japanese modern literature, uses its specific markup language for digitalizing texts that is called Aozora Chuki (Open Air Annotations). Particular typesetting and composing of Japanese books have forced volunteer archivists to invent and manage their own markup organized in Japanese language, which enable and encourage those who understand only Japanese to join in the digital archive project easily. The development of making e-texts in Aozora Bunko for 20 years has served to a new plan of JIS kanji code, novel applications of text viewers for Japanese e-books and universal usage of digital texts, for example, automatic reading aloud or braille translation. Nowadays, these marking-up annotations contribute to professional and amateur authors who write their works digitally in Japanese language.

This paper makes a brief report on the history of its own digitalization rule in Aozora Bunko, which includes successes and problems in Japanese typesetting and marking-up: ruby letters, dots alongside words, character codes, code unification, accent mark separation, interlinear notes and so on. The main purpose of these markups is to make a digital reproduction of textual information in Japanese books via simple text files. This report will offer a new topic or viewpoint for discussion among text markup languages.

Text Encoding in Large-Scale Corpus Construction: the Case of BCCWJ and CHJ
Presenter**: Yoko Mabuchi**

We at the National Institute for Japanese Language and Linguistics have constructed two kinds of large-scale written Japanese corpora and have them open to the public. One is a Contemporary Japanese corpus which includes more than one hundred million words. The other is a historical Japanese corpus of about 15 million words from various documents of a wide range of periods which includes from “Man’yōshū” (the oldest

[1] Kyoto Tachibana University / Aozora-bunko, Japan
[2] National Institutes for the Humanities, Japan
[3] International Institute for Digital Humanities, Japan

existing collection of Japanese poetry) compiled around the 7th-8th century to modern magazines of early 20th century.

Many users, such as researchers of the Japanese language, students studying Japanese linguistics, foreigners learning Japanese as a second language, use these corpora through the online search application called “Chunagon”.

A feature of these corpora is that it has detailed word information. The data before morphological analysis is encoded in the XML format, and the text format is designed so as to be compatible with TEI. I am involved in the development of text encoding for both corpora. In this presentation, I report the following three points: 1) the search for the framework of uniformly marking up documents of various forms and types and structuralization appropriate for Japanese language research (the main purpose of use of the corpus), 2) problems in encoding documents of historical Japanese, 3) difficulties experienced when we tried to structure special printing formats. Then I suggest topics for the panel discussion.

A Use Case of Born Digital Modern Japanese Texts

Presenter: **Kiyonori Nagasaki**

SAT Daizokyo text database project (led by Prof. Masahiro Shimoda, henceforth, SAT project) released some modern Japanese translations encoded by TEI, when the SAT project launched its new Web service called SAT2018 in the end of March, 2018. It was not difficult to encode the Japanese texts by TEI in this case because it doesn’t require any special treatment due to its purpose and natures of the texts.

The encoding aimed to markup proper names, titles, sentences, and dialogues to facilitate understanding and processing of the texts. The markup itself was so simple. This case focused on handling, linking and visualization of the TEI-encoded texts with jQuery.

Displaying a text body, the viewing system, that is, SAT2018 change the color of place name with red and person name with blue. Chapter titles marked up in the text are extracted and listed in a floating window. User can move to arbitrary chapter by clicking a chapter title. (Fig.1)

The sentence elements <s> are used to align a modern Japanese text with a classical Chinese text which was translated into Japanese. Each <s> is connected with line number of the classical Chinese text via xml:id added to <s>. (Fig.2)

The alignment was edited by a collaboration system on SAT2018. The system allows to align both texts by mouse manipulation. (Fig.3) As the Japanese texts include <s> with xml:id, users can select a sentence only by one click. But, as the Chinese text are not yet separated by sentence. it is necessary to select a part of text by dragging.

This case would be useful for marking up similar kind of texts.

Figure 1: Colored texts with its index

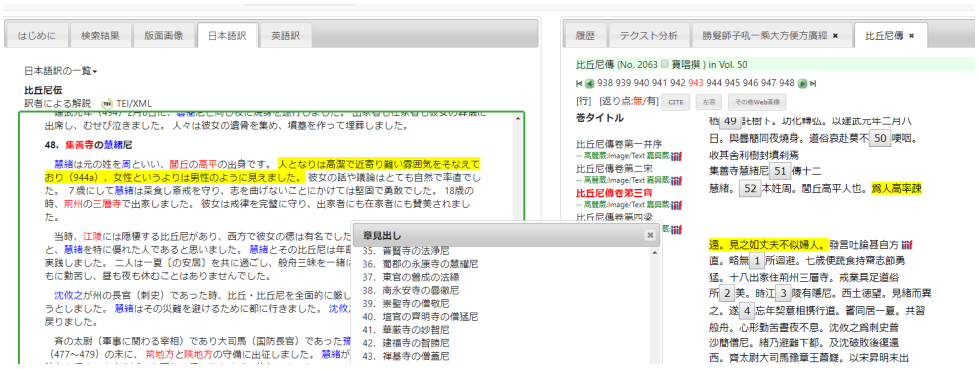


Figure 2: Alignment between Japanese and Chinese texts

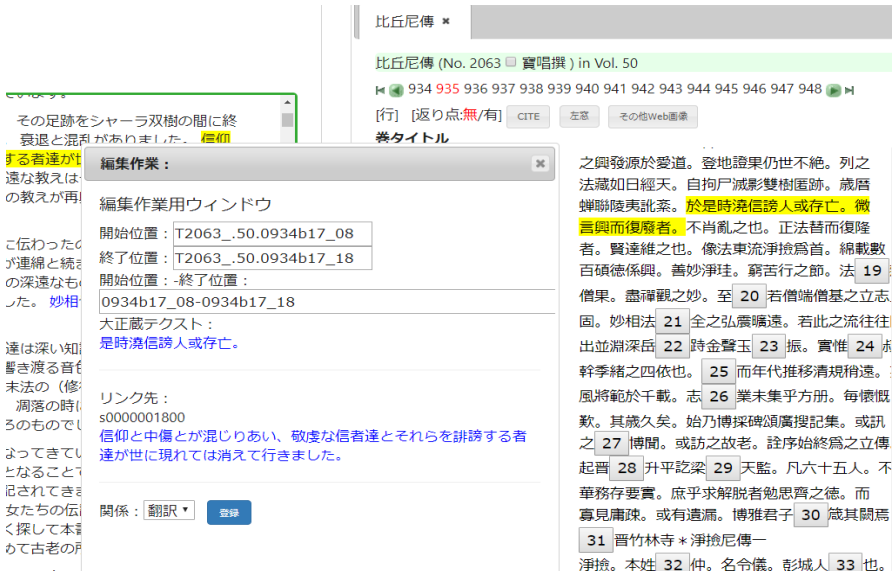


Figure3: Editing alignment on SAT201

[Day 5 (Thu) 14:55-16:25] LP Session X (Room A1)
Lexicography and Language

The TEI in Real-Time Lexicography: *The Digital Dictionary of Buddhism* and *CJKV-E Dictionary* after 32 Years

A. Charles Muller[1]

While there is a plethora of TEI-based projects that represent efforts to make sense of, archive, organize, or analyze some form of pre-existent form (usually paper) of data, there are far fewer that are "born digital" and even fewer that are "born digital in TEI matrix", and yet even fewer that have gone on to become major online resources, delivering the data directly from (near-)TEI format onto the web through XSLT-generated web pages. Thus, the Digital Dictionary of Buddhism [DDB] (http://www.buddhism-dict.net/ddb/) and its companion CJKV-E Dictionary (http://www.buddhism-dict.net/dealt) (both online since 1995) together constitute a somewhat unique and powerful example of the possibilities of TEI-based text encoding, one that has served as the basis for what is arguably one of the largest and most successful examples of scholarly-based crowdsourcing on the entire web—and probably the one with the longest history. The longevity and the ongoing steady growth of the project are in themselves a testament to the original basic aims of the TEI, of providing a long-term, flexible, and sustainable means of text encoding, allowing its users to avoid the traps and limitations of propriety software. This presentation will start off by offering a brief overview of the history and development of the project, including its major developmental transitions. We will then focus on the aspects of TEI in particular that have played the most important roles in allowing the attainment of vibrant crowd-sourced sustainability.

TEI Metadata for Japanese Materials in the Cambridge University Library and How to Apply TEI for Higher Education

Zengxian Li[2]

This paper discusses TEI metadata for the early Japanese books in the Cambridge University Library (CUL) and application of TEI for higher education. The author's affiliation, Art Research Center (ARC), is an institute with rich experiences on digitizing Japanese traditional cultural materials. Since its foundation in 1996, it has been participating with a considerable number of institutes all cross the world. Since 2015, the author has engaged in digitizing its collection, putting a significant number of the books online through the CUL's official depository, the Cambridge Digital Library (CDL).

The CDL requires the books' metadata to be prepared in TEI format, which the author was not familiar with. Yet, the kind help from the Digital Contents Unit in the CUL enabled us to publish over 80 titles, including 200 volumes of early Japanese books to the public. The author finds learning how to make metadata in a specific format valuable experience, and the experience also gives him some ideas on how to apply TEI for the higher education.

[1] The University of Tokyo, Japan
[2] Ritsumeikan University, Japan

In the field of classical Japanese literature, researchers and students conduct research based on ancient text which often has textual variations. In such a case, research starts with comparing those texts and manually listing their differences. However, the author found out the fact that by applying TEI structure to the text, one can leave the manual task to the machine. Moreover, he also found out that by using well-known literary works, such as the Kokin-wakashu (the oldest imperial anthology of Japanese poems), one can help students' better understanding about how the TEI could be involved with classic Japanese literary works, which eventually develops their interests in both TEI and the literary works.

TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

Laurent Romary[1], Toma Tasovac[2]

Achieving consistent encoding within a given community of practice has been a recurrent issue for the TEI Guidelines. The topic is of particular importance for lexical data if we think of the potential wealth of content we could gain from pooling together the information available in the variety of highly structured, historical and contemporary lexical resources. Still, the encoding possibilities offered by the Dictionaries Chapter in the Guidelines are too numerous and too flexible to guarantee sufficient interoperability and a coherent model for searching, visualising or enriching multiple lexical resources.

Following the spirit of TEI Analytics [Zillig, 2009], developed in the context of the MONK project, TEI Lex-0 aims at establishing a target format to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such [Ermolaev and Tasovac, 2012] and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers. The format itself should not necessarily be one which is used for editing or managing individual resources, but one to which they can be univocally transformed to be queried, visualised, or mined in a uniform way. We are also aiming to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF (Lexical Markup Framework) standard so that coherent design guidelines can be provided to the community (cf. [Romary, 2015]).

The paper will provide an overview of the various domains covered by TEI Lex-0 and the main decisions that were taken over the last 18 months: constraining the general structure of a lexical entry; offering mechanisms to overcome the limits of <entry> when used in retro-digitized dictionaries (by allowing, for instance, <pc> and <lbl> as children of <entry>); systematizing the representation of morpho-syntactic information [Bański et al., 2017]; providing a strict <sense>-based encoding of sense-related information; deprecating <hom>; dealing with internal and external references in dictionary entries, providing more advanced encodings of etymology (see submission by Bowers, Herold and Romary); as well as defining technical constraints on the systematic use of @xml:id at different levels of the dictionary microstructure. The activity of the group has already lead to changes in the Guidelines in response to specific GitHub tickets[3].

Acknowledgements
The initiative started in the context of the COST Action European Network for e-Lexicography (ENEL) and is now supported by DARIAH, as part of the activities of the

[1] Inria (team ALMAnaCH), France and DARIAH
[2] Belgrade Center for Digital Humanities, Serbia
[3] See for instance https://github.com/TEIC/TEI/issues/1702 (model.entryPart.top for <pc> and <lbl>), https://github.com/TEIC/TEI/issues/1688 (add <form> to att.typed) or https://github.com/TEIC/TEI/issues/1734 (make hyph/stress/syll members of att.notated)

Lexical Resources working group[1], as well as the H2020-funded project European Lexicographic Infrastructure (ELEXIS)[2].

References

Bański, Piotr, Jack Bowers and Tomaz Erjavec (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, Sep Leiden, Netherlands. 〈hal-01757108〉

Ermolaev, Natalia, and Toma Tasovac (2012) "Building a Lexicographic Infrastructure for Serbian Digital Libraries." *Libraries in the Digital Age (LIDA) Proceedings* 12, no. 0 (June 12). http://ozk.unizd.hr/proceedings/index.php/lida/article/view/55.

ISO 24613:2008 *Language resource management - Lexical markup framework (LMF)*, currently revised as a multipart standards with Part 1: core model, Part 2: Machine Readable Dictionaries, Part 3: Etymology, Part 4: TEI serialisation

Romary, Laurent (2015). TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics,* 30 (1), 〈http://www.jlcl.org〉. 〈hal-00762664v4〉

Zillig, Brian (2009) "TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text Analysis." *Literary and Linguistic Computing* 24 (2009): 187–192. https://doi.org/10.1093/llc/fqp005

[1] https://www.dariah.eu/activities/working-groups/lexical-resources/

[2] https://www.cordis.europa.eu/project/rcn/213379_en.html

[Day 5 (Thu) 14:55-16:25] Panel Session IV (Room A2)

Whither TEI Publisher?

Magdalena Turska[1], Wolfgang Meier[1]

In this slightly different take on panel session we'd like to tap into collective wisdom of the TEI community and discuss the possible development avenues for the TEI Publisher and TEI Processing Model.

TEI Publisher is an open source application generator for eXist-db platform implementing the TEI Processing model. TEI Publisher allows to publish digital editions from data encoded in TEI (as well as other XML vocabularies e.g. DocBook) with full text search, navigation, page- or -division based views and other typical functionality provided out of the box, as well as PDF and ePub formats beside HTML. While it has been successfully used in numerous research and publishing projects the authors contributed to, there are still some challenges when research teams are working on customizing the TEI Publisher on their own.

While users typically become quickly fluent in tweaking models in the ODD, customizations of the layout or application CSS styles are not as straightforward. We would like to provide a wider range of sensible default layouts and functional components and thus need the input of the community concerning use cases and specific requirements. Aiming to make TEI Publisher easier to use for researchers, even inexperienced in web technologies such as JavaScript or XQuery, we'd appreciate an open discussion about community needs in that area as well as what our current and prospective users see as major difficulties.

First part of the session, presented by Magdalena Turska will briefly familiarize the audience with the current state of the TEI Publisher (now at version 3.0) and range of projects that have already adopted it, stressing the common factors and demonstrating areas where customization was necessary.

In the second part, Wolfgang Meier will report on experiences from supporting TEI Publisher based projects, main challenges identified so far and sketch ideas to facilitate adoption for less technical users and broader range of projects.

The last part we'd like to treat as brainstorming session and public discussion, thus we invite everyone interested as our collective 3rd panel speaker.

1. TEI Publisher: current state of affairs (Magdalena Turska)
2. Challenges of TEI Publisher customization (Wolfgang Meier)
3. Discussion (TEI Community)

[1] eXist Solutions, Germany