



TEI 2022 Conference Book

12 - 16 September 2022
Newcastle University
Newcastle Upon Tyne, UK

Edited by James Cummings - 2022
Version 1.1

Table of Contents

[Table of Contents](#)

[Welcome From the Local Organisers](#)

[Welcome From the Chair of the Programme Committee](#)

[The Programme Committee](#)

[TEI2022 Joining Instructions](#)

[Venue Location](#)

[Accommodation](#)

[Conference Programme](#)

[Workshops:](#)

[Presenters](#)

[Food and Refreshments](#)

[Covid19](#)

[WiFi and Computer Access](#)

[Code of Conduct](#)

[Quiet Space](#)

[Dress Code](#)

[Armstrong Building Maps](#)

[Previous TEI Consortium Conferences](#)

[TEI2022 Conference Programme](#)

[The overall schedule](#)

[Keynotes Lectures](#)

[Opening Keynote: Constance Crompton, "Situated, Partial, Common, Shared: TEI Data as Capta"](#)

[Closing Keynote: Emmanuel Ngue Um, "Tone as "Noiseless Data": Insight from Niger-Congo Tone Languages'](#)

[Workshops](#)

[Workshop 1: From a collection of documents to a published edition : how to use an end-to-end publication pipeline \[Full Day\]](#)

[Workshop 2: Creating Digital Editions with FairCopy \[Half Day, Afternoon\]](#)

[Workshop 3: A short introduction to Schematron \[Half Day, Afternoon\]](#)

[Workshop 4: Building TEI-powered websites with static site technology. A hands on exploration of the publishing toolkit of the Scholarly Editing Journal \[Half Day, Morning\]](#)

[Workshop 5: Introduction to XProc \[Half Day, Morning\]](#)

[Workshop 6: Engaging TEI Editors Through LEAF-Writer \[Half Day, Morning\]](#)

[SIG Meetings](#)

[Conference Sessions – Wednesday 14 September 2022](#)

[Session 1A – Short Papers – 09:30 - 11:00](#)

[Short Paper: Standoff-Tools. Generic services for building automatic annotation pipelines around existing tools for plain text analysis](#)

[Short Paper: TEI Automatic Enriched List of Names \(TAELN\): An XQuery-based Open Source Solution for the Automatic Creation of Indexes from TEI and RDF Data](#)

[Short Paper: manuForma – A Web Tool for Cataloging Manuscript Data](#)

[Session 1B – Long Papers – 09:30 – 11:00](#)

[Long Paper: Texts All the Way Down: The Intertextual Networks Project](#)

[Long Paper: Revising Sex and Gender in the TEI Guidelines](#)

[Long Paper: Where is the Spanish in the TEI?: Insights on a Bilingual Community Survey](#)

[Session 2A – Long Papers – 11:30 – 13:00](#)

[Long Paper: Revision, Negation, and Incompleteness in Melville's Billy Budd Manuscript](#)

[Long Paper: “Un mar de sentimientos”. Sentiment analysis of TEI encoded Spanish periodicals using machine learning](#)

[Session 2B – Long Papers – 11:30 – 13:00](#)

[Long Paper: TEICollator: a semi-automatic TEI to TEI workflow](#)

[Long Paper: Back to analog: the added value of printing TEI editions](#)

[Long Paper: Encoding sonic devices: what is it good for?](#)

[Session 3A – Long Papers – 14:30 – 16:00](#)

[Long Paper: Vocabularium Bruxellense. Towards Quantitative Analysis of Medieval Lexicography](#)

[Long Paper: ISO MAF reloaded: new TEI serialization for an old ISO standard](#)

[Long Paper: TEI Modelling of the Lexicographic Data in the DARIAH-PL Project](#)

[Session 3B – Panel: Notes from the DEPCHA Field and Beyond: TEI/XML/RDF for Accounting Records – 14:30 – 16:00](#)

[Posters Slam and Session – 16:30 – 18:00](#)

[Poster: The QhoD project: A resource on Habsburg-Ottoman diplomatic exchange](#)

[Poster: Building a digital infrastructure for the edition and analysis of historical travelogues](#)

[Poster: TEI and Scholarly Digital Editions: how to make philological data easier to retrieve and elaborate](#)

[Poster: Between Data and Interface, Building a Digital Library for Spanish Chapbooks with TEI-Publisher](#)

[Poster: oXbytei and oXbytao. A Stack of Configurable oXygen Frameworks](#)

[Poster: Automatic Validation, Packaging and Deployment of TEI Documents. What Continuous Integration can do for us](#)

[Poster: Adapting TEI for Braille](#)

[Poster: Okinawan Lexicography in TEI: Challenges for Multiple Writing Systems](#)

[Poster: Text as Object: Encoding the data for 3D annotation in TEI](#)

[Poster: Building Interfaces for East Asian/Japanese TEI data](#)

[Poster: Explainable Supervised Models for Bias Mitigation in Hate Speech Detection: African American English](#)

[Poster: A TEI/IIIF Structure for Adding Palaeographic Examples to Catalogue Entries](#)

[Poster: From facsimile to online representation. The Centre for Digital Editions in Darmstadt. An Introduction](#)

[Poster: From Oxgarage to TEIGarage and MEIGarage](#)

[Poster: Towards a digital documentary edition of CCCC41: The TEI and Marginalia-Bearing Manuscripts](#)

[Poster: Transatlantic Networks - a Pilot: mapping the correspondence of David Bailie Warden \(1772-1845\)](#)

[Conference Sessions – Thursday 15 September 2022](#)

[Session 4A – Short Papers – 09:30 - 11:00](#)

[Short Paper: TEI and the Re-Encoding of Born-Digital and Multi-Format Texts](#)

[Short Paper: Capturing the Thread Structure: A Modification of CMC-Core to Account for Characteristics of Online Forums](#)

[Short Paper: Publishing the grammateus research output with the TEI : how our scholarly texts become data](#)

[Short Paper: Handwritten Text Recognition for heterogeneous collections? The Use Case Gruß & Kuss](#)

[Session 4B – Long Papers – 09:30 - 11:00](#)

[Long Paper: From TEI Personography to IPIF data](#)

[Long Paper: TEI as Data: Escaping the Visualization Trap](#)

[Long Paper: LINC'S' Linked Workflow: Creating CIDOC-CRM from TEI](#)

[Session 5A – Long Papers – 11:30 - 13:00](#)

[Long Paper: Evolving Hands: HTR and TEI Workflows for cultural institutions](#)

[Long Paper: Between automatic and manual encoding: towards a generic TEI model for historical prints and manuscripts](#)

[Long Paper: Dehmel Digital: Pipelines, text as data, and editorial interventions at the distance](#)

[Session 5B – Panel: Manuscript catalogues as data for research – 11:30 - 13:00](#)

[Session 6A – An Interview With ... Lou Burnard – 14:30 - 16:00](#)

[TEI Consortium Annual General Meeting – 16:30 - 18:00](#)

[Conference Sessions – Friday 16 September 2022](#)

[Session 7A – Short Papers – 09:30 - 11:00](#)

[Short Paper: Encoding Complex Structures: The Case of a Gospel Spanish Chapbook](#)

[Short Paper: Annotating a historical manuscript as a linguistic resource](#)

[Short Paper: How to Represent Topic Models in Digital Scholarly Editions](#)

[Short Paper: Analyzing the Catalogue of Heroines through Text Encoding](#)

[Session 7B – Long Papers – 09:30 - 11:00](#)

[Long Paper: Is it still data? Scholarly Editing of Text from Early Born-Digital Heritage](#)

[Long Paper: Using Citation Structures](#)

[Long Paper: Text between data and metadata: An examination of input types and usage of TEI encoded texts](#)

[Session 8A – Long Papers – 11:30 - 13:00](#)

[Long Paper: Codex as Corpus : Using TEI to unlock a 14th-century collection of Old French short texts](#)

[Long Paper: atop: another TEI ODD processor](#)

[Session 8B – Demonstrations – 11:30 - 13:00](#)

[Demonstration: Transcribing Primary Sources using FairCopy and IIIF](#)

[Demonstration: Adapting CETELcean for static site building with React and Gatsby](#)

[Demonstration: Spec Translator: Enabling translation of TEI Specifications](#)

[Demonstration: LEAF-Writer: a TEI + RDF online XML editor](#)

[Virtual Poster Session on Gather.Town – Thursday 22 September 2022](#)

[Virtual Poster: From Archives to TEI Publisher: Digital Edition of German Work Regulations in the Project 'Non-state Law of the Economy'](#)

[Virtual Poster: Feature structures for character social variable annotation and an application to Alsatian theater](#)

[Virtual Poster: Multilingualism and multiscryptism in TEI publishing: DH2022](#)

[Virtual Poster: Celebrating Deviation: Encoding Variant Japanese Phonetic Characters known as Hentaigana](#)

[Virtual Poster: Theoretical and practical challenges of automatically identifying and encoding alliteration in texts written in Italian](#)

[Closing Note](#)

Welcome From the Local Organisers

We would like to welcome you to the TEI2022 conference. The organisation of a conference of any sort is filled with various stresses and anxieties as you balance the economics of the costs of running the conference with the desire to provide attendees the most enjoyable opportunity for academic interchange of ideas possible. We thank our sponsors for the help they have given us in running the conference. In 2022 events are especially challenging as the return to in-person activities after the public health restrictions of the Covid19 pandemic have been lifted. While the number of attendees might be slightly reduced from pre-pandemic years it is hoped that we can all meet in a healthy and safe manner. While a fully hybrid conference was not feasible with the institutional setup we currently have, we are using Newcastle University's ReCap system (usually for students who have missed lectures) to record the screens and audio of most of the parallel sessions and keynotes. We hope to make those openly available shortly after the conference. The venue for the conference is the lovely Armstrong Building on the Newcastle University campus, including our breaks and receptions in the grand King's Hall. This conference book, filled with the abstracts of the presentations and other useful information has been made mostly by copying-and-pasting the final doc/docx/odt files uploaded by the participants. Although a manual process this gave more direct awareness of any immediate issues. While some light editing for formatting has been done, this has not been done rigorously and various inconsistencies remain. *Mea culpa*.

The conference comes immediately following the sad news of the death of Queen Elizabeth II and accession of King Charles III. Vice-Chancellor and President of Newcastle University, Professor Chris Day who is also the Deputy Lieutenant of Tyne and Wear said: "The University community joins the Royal Family and the nation in mourning Her Majesty Queen Elizabeth II. In keeping with Government guidelines and advice from the University, there will be no changes to the conference programme.

We hope that your conference is academically and socially engaging. If you have any difficulties during the conference, do not hesitate to seek out the help of the local organisers and student assistants.

James Cummings,
Adam Mearns,
Tiago Sousa Garcia

TEI2022 Local Organisers
September 2022

Welcome From the Chair of the Programme Committee

After three years, the 21st annual Conference and (22nd) Members' Meeting of the Text Encoding Initiative Consortium (TEI) can finally take place in person again. The conference with the theme "Text as data" is organised by Newcastle University.

The past decade has seen a huge increase of data produced by (social) media platforms, digital literary outputs, and various mass digitization efforts of cultural heritage and administrative records. Though these vast data collections hold enormous potential for diverse research, collecting and analysing text-based data also presents unique challenges that need to be addressed. The increasing quantity of the textual data coincides with its improved availability and accessibility, but also the continuously progressing development of data models, tools, text-mining, and machine-learning techniques. The TEI community is working at the intersection of many of these areas.

If we want the computer to "understand" a text we must either mark textual phenomena or instruct a computer to identify them. In their acclaimed work "The Shape of Data in the Digital Humanities" from 2018, Julia Flanders and Fotis Jannidis refer to this as "a choice between an algorithmic approach [...] or what we might call a "metatextual" approach, in which information is added to the text in some explicit form that enables it to be processed intelligently".

Inspired by these considerations, this year's contributions are covering a broad scope of disciplines, topics and methods such as manuscript studies, literary studies, linguistics, lexicography, history, digital scholarly editing, text processing, publishing, handwritten text recognition, linked open data, semantic web technologies, and many more. Posters, demonstrations, and workshops were evaluated by at least two reviewers from the program committee; panels as well as long and short paper proposals by at least three reviewers. The authors had the opportunity to follow up on the review results and extend their abstracts for the final version of their contributions which you find in this book of abstracts.

In total 11 short papers, 22 long papers, 2 panels, 16 posters, 5 virtual posters, 4 demonstrations, and 6 workshops were accepted. The new "An interview with ..." format, which highlights long-time and expert contributions to the TEI, is introduced with Lou Burnard as the first guest. Additionally, Constance Crompton is presenting the opening keynote "Situated, Partial, Common, Shared: TEI Data as Capta", and Emmanuel Ngue Um will bring the conference to close with his keynote "Tone as 'Noiseless Data': Insight from Niger-Congo Tone Languages".

We want to thank the sponsors for their financial support, all contributors for their wide-ranging and diverse submissions, the program committee for their diligent efforts in ensuring the high quality of the conference, and above all the local organisers James Cummings, Adam Mearns, and Tiago Sousa Garcia for their tireless work to make a successful and memorable conference possible.

Martina Scholger
Chair of the Programme Committee
September 2022

The Programme Committee

The content of the programme itself is determined by the Programme Committee:

- Bernhard Bauer (University of Graz, Austria)
- Syd Bauman (Northeastern University, USA)
- Elisa Beshero-Bondar (Penn State Behrend, USA)
- Elli Bleeker (Huygens Institute for the History of the Netherlands, Netherlands)
- Meaghan Brown (NEH Research Division, USA)
- Gabriel Calarco, (CONICET, Argentina)
- Hugh Cayless (Duke University, USA)
- James Cummings (Newcastle University, United Kingdom)
- Gimena del Rio Riande (CONICET, Argentina)
- Gustavo Fernández Riva (University of Heidelberg, Germany)
- Christiane Fritze (Vienna City Library, Austria)
- Ulrike Henny-Krahmer (University of Rostock, Germany)
- Martin Holmes (University of Victoria, Canada)
- Diane Jakacki (Bucknell University, USA)
- Dario Kampkaspar (Universitäts- und Landesbibliothek Darmstadt, Germany)
- Johannes Kepper (Paderborn University, Germany)
- Frederike Neuber (Berlin-Brandenburgische Akademie der Wissenschaften, Germany)
- Trisha O'Connor (University of Oxford, United Kingdom)
- Laurent Romary (INRIA, France)
- Martina Scholger (University of Graz, Austria) – Chair
- Peter Stadler (Paderborn University, Germany)
- Kathryn Tomasek (Wheaton College Massachusetts, USA)
- Yifan Wang (International Institute for Digital Humanities, Japan)

TEI2022 Joining Instructions

These are the joining instructions for the TEI2022 conference, happening at Newcastle University on the 12-16 September 2022. Thank you for registering for the conference and we are looking forward to welcoming you.

If you have any questions, don't hesitate to contact tei2022@ncl.ac.uk.

Venue Location

The [venue](#) for the TEI2022 conference is the [Armstrong Building on the Newcastle University campus](#). **Building Work:** there is currently building work with no firm completion date happening to the front of the building on Queen Victoria Road. This means the entrance you are recommended to use is via [King's Quad](#). To enter King's Quad go through the archway (at [What3Words.com/Strut.Vote.Code](https://www.what3words.com/Strut.Vote.Code)) and pass the statue of Martin Luther King to the registration desk. (If you are curious about the background to this, see this Special Collections Exhibit on [Martin Luther King at Newcastle University](#).) We have produced [a PDF of floor plans of the Armstrong Building](#) with conference rooms highlighted.



If arriving at Newcastle Airport, the easiest way into town is by the Metro system. The stop closest to the venue is at 'Haymarket', and the centre of the city is at the next stop 'Monument'. Once in Newcastle, most people find it walkable.

For more information see:

- <https://conferences.ncl.ac.uk/tei2022/about/travelinformation>
- <https://conferences.ncl.ac.uk/tei2022/about/aboutnewcastle>
- <https://conferences.ncl.ac.uk/tei2022/about/aboutnewcastleuniversity>
- <https://conferences.ncl.ac.uk/tei2022/about/venueinformation>
- <https://conferences.ncl.ac.uk/tei2022/about/venueinformation/TEI2022-floorplans.pdf>

Accommodation

Accommodation booking can be made via [this hotel booking portal](#), or directly with the hotels. If you are finding it difficult to find a hotel for Sunday 11 September 2022, this is because it is the day of the very popular [Great North Run](#) Half-Marathon. See our [Accommodation Page](#) for more information.

Conference Programme

The full programme of the conference is available at:

<https://www.conftool.pro/tei2022/sessions.php>. The general structure of the conference is:

Times / Days	Monday 12 September 2022	Tuesday 13 September 2022	Wednesday 14 September 2022	Thursday 15 September 2022	Friday 16 September 2022
09:00 - 09:30	Refreshments in King's Hall				
09:30 - 11:00	Workshops	Workshops	Session 1	Session 4	Session 7
11:00 - 11:30	Break in King's Hall				
11:30 - 13:00	Workshops	Workshops	Session 2	Session 5	Session 8
13:00 - 14:30	Buffet Lunch in King's Hall				
14:30 - 16:00	Workshops	SIG Meetings	Session 3	Session 6	Closing Keynote in ARMB.2.98
16:00 - 16:30	Break in King's Hall				Reception in King's Hall
16:30 - 18:00	Workshops	SIG Meetings	Poster Session in King's Hall	TEI-C AGM in ARMB.2.98	
18:15 - 19:30		Opening Keynote in ARMB.2.98			
19:30 - 21:00		Reception in King's Hall			

Workshops:

If you signed up for a workshop, you should show up at the room and time shown in the conference schedule. If you now cannot make it, please consider letting the workshop organiser know. Your workshop organiser has been given a list of those who signed up and may contact you with information about the workshop. You are expected to bring your own laptop to the workshop, the conference is not supplying computers for workshops. If you did not sign up for a workshop and now would like to attend one, either email the workshop organiser or show up at the appointed date/time and ask if you can join it. (Workshop organisers are under no obligation to accept walk-ins.)

Presenters

There is information about presenting at TEI2022 at:

<https://conferences.ncl.ac.uk/tei2022/about/presenterinformation>

This includes information on the [presenting computers](#), [recording](#), [timings](#), [chairing](#), [poster session](#), and [virtual poster session](#) (at which all physical poster presenters are also invited to present).

Food and Refreshments

There are refreshment breaks (providing coffee, tea, and snacks) at 9am, 11am, and 4pm each day. Lunch at 1pm each day is a standing buffet with a selection of wraps and other foods. Dietary preferences that you noted while registering have been passed to the university catering team. There are two receptions after the opening and closing keynotes which will have drinks and nibbles.

In a post-pandemic world, the local organisers took the decision not to host a conference banquet. Instead we will be assisting self-arranging groups looking for those to join them to do so in a more informal manner. If you are interested in making a registration for a group of people, or want to join a group looking for more people, then please add the appropriate details to our [TEI2022 Sign-up Groups for Evening Meals](#).

Covid19

While the UK currently has no Covid-related restrictions at all, we wish to make the TEI2022 conference as safe as an in-person conference can be. We strongly suggest participants test themselves before travelling to the conference, wear a mask where feasible during the conference, and test themselves when they return home. If you test positive, we'd request that you not attend the conference. Tests are available from pharmacies in the UK such as Boots (on the other side of Haymarket Metro Station, or in the Eldon Square Shopping Centre). Disposable masks will be available free from the registration desk.

WiFi and Computer Access

The conference venue has good WiFi coverage and the EduRoam network should be preferred if you use it at your institution. (We recommend connecting to EduRoam at your institution at least once prior to coming to the conference.) If your institution does not use EduRoam there is 'Guest WiFi' which is run by Sky's ['The Cloud'](#) service. You will need to register for an account with them to use that WiFi but it is otherwise free of charge.

All presenting computers will have logins posted next to them which will work on presenting computers in the conference venue.

Code of Conduct

Everyone should be able to enjoy the TEI-C Conference, and so there is absolutely no place at the TEI-C Conference for harassment or intimidation of any sort. Please see the TEI-C Conference Code of Conduct at: <https://conferences.ncl.ac.uk/tei2022/about/conduct/>.

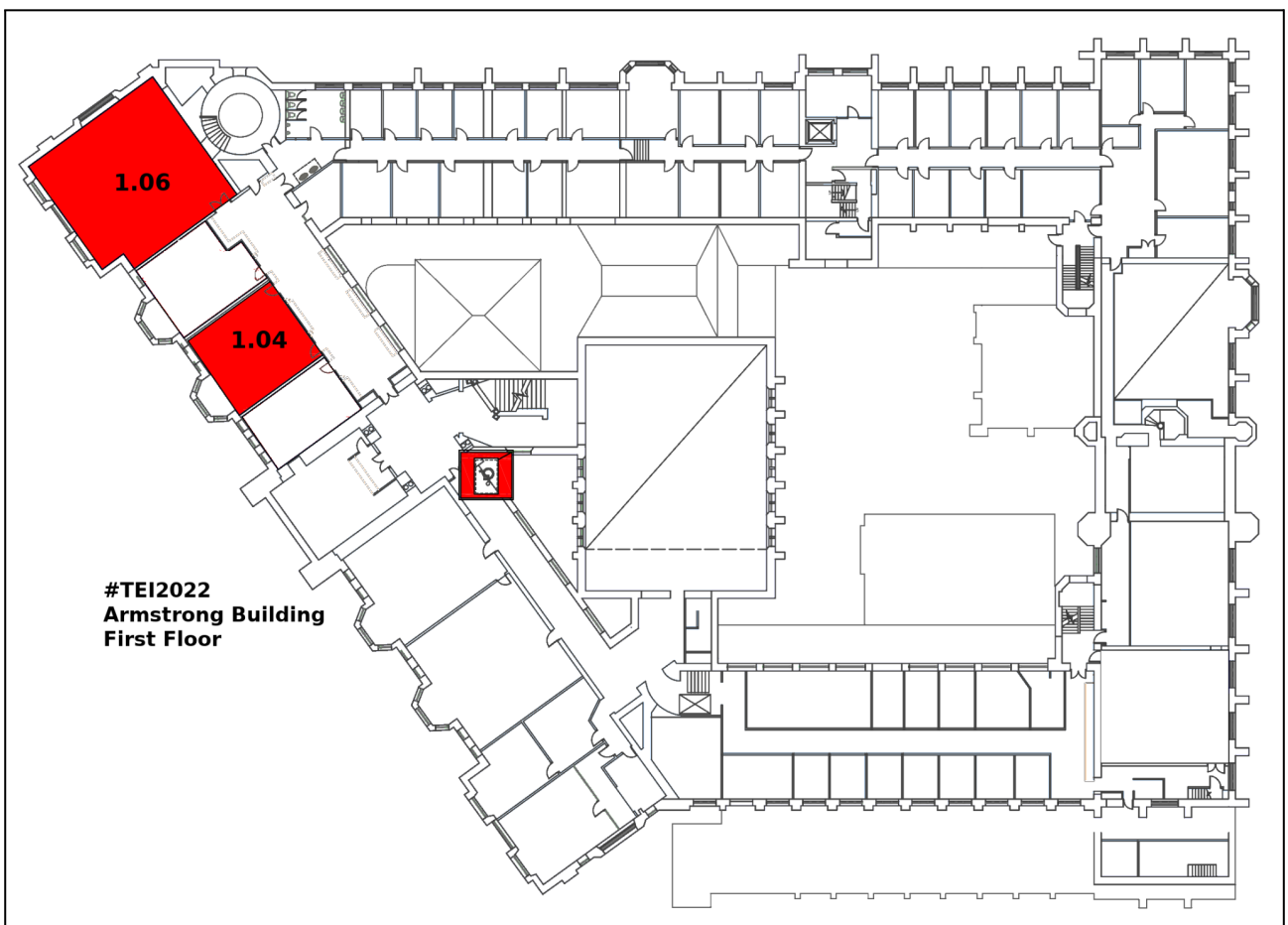
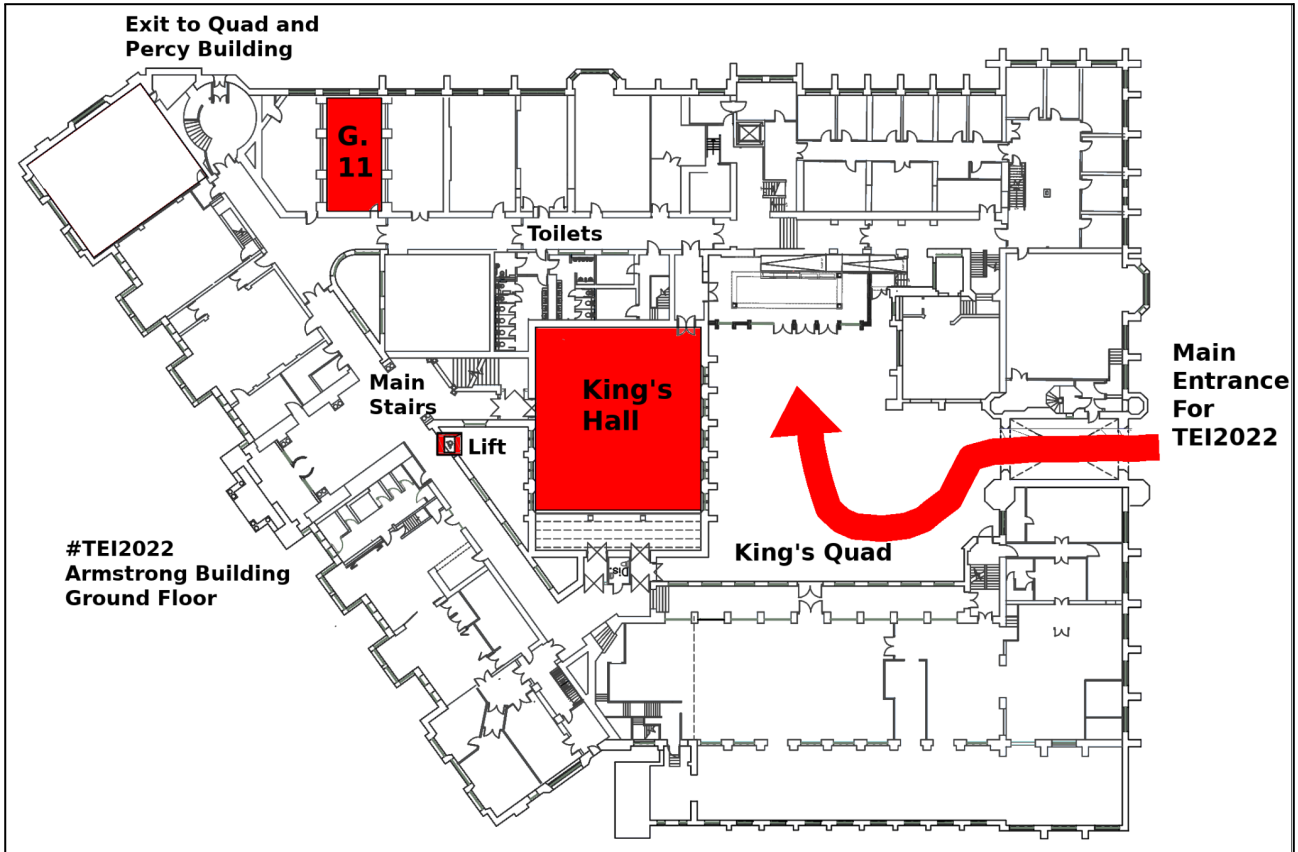
Quiet Space

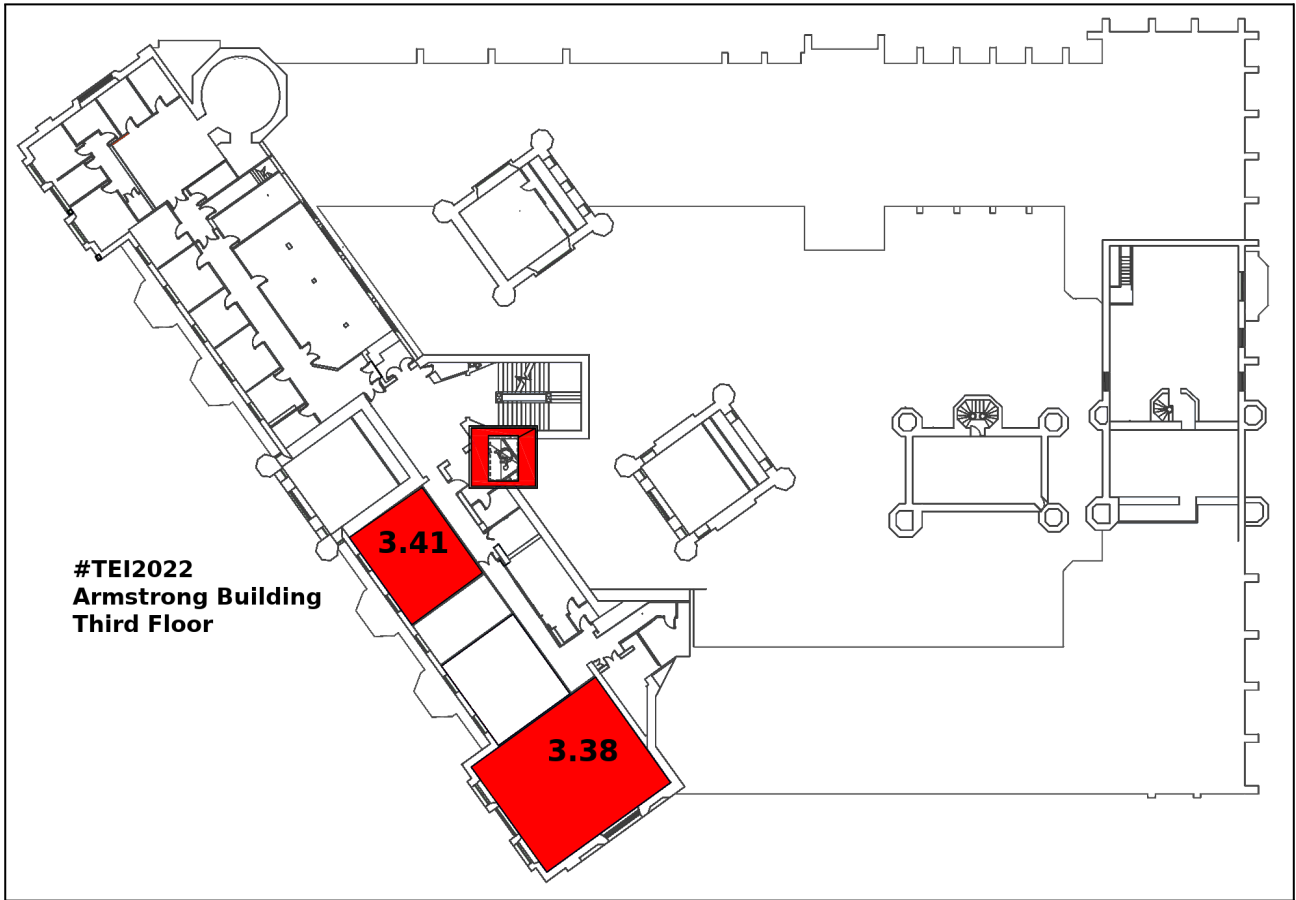
If there are those who need a space away from the buzz of the rest of the conference, we have reserved room [ARMB.G.11](#) as a room with no other specific purpose. It is not legally a creche for childcare (because that has specific legal requirements) but you could choose to look after children there. Although we've called it a Quiet Space, it also has a piano and is usually a music seminar room (so calling it 'quiet' may not be accurate).

Dress Code

As we return to in-person conferences, some may wish to dress more formally, others for comfort. There is no dress code, wear what you like.

Armstrong Building Maps





Previous TEI Consortium Conferences

The TEI's conference and annual general meeting is the gathering point for the TEI community. It offers an opportunity to meet with colleagues, learn about new projects, share research, and find out about new developments in the TEI. Originally structured around the annual TEI business meeting and election, the event has more recently expanded to include a full academic conference program with peer-reviewed papers, posters, tool demonstrations, and meetings of the TEI special interest groups. The conference and members' meeting is also an excellent opportunity for businesses and other organisations in the Digital Humanities community to reach active researchers through sponsorship and other support activities.

Past and future meetings are listed below. Conference websites of meetings from 2009 and later are also archived in our [Members' Meetings Website Archive](#).

- [Newcastle, UK, 2022](#)
The 21st annual conference and annual general meeting is being held September 12-16, 2022.
- [Next Gen TEI, virtual, 2021](#)
The 20th annual conference and annual general meeting was held October 25-27, 2021.
- 2020: The 2020 conference and annual general meeting, to be held in Lincoln, Nebraska, was replaced with an online meeting in November.
- [Graz, Austria, 2019](#)
The 19th annual conference and annual general meeting was held September 16-20, 2019.
- [Tokyo, Japan, 2018](#)
The 18th annual conference and annual general meeting was held September 9-13, 2018.
- [Victoria, British Columbia, 2017](#)
The 17th annual conference and annual general meeting was held November 11-15, 2017.
- [Vienna, Austria, 2016](#)
The 16th annual conference and annual general meeting was held September 26-30, 2016.
- [Lyon, France, 2015](#)
The 15th annual conference and annual general meeting was held October 26-31, 2015.
- [Evanston, Illinois USA, 2014](#)
The 14th annual conference and annual general meeting was held October 22-24, 2014, hosted by Northwestern University.
- [Rome, Italy, 2013](#)
The 13th annual conference and annual general meeting was held October 2-5, 2013, hosted by Università La Sapienza.
- [College Station, Texas, 2012](#)
The 12th annual conference and annual general meeting was held November 7-10, 2012, hosted by Texas A & M University.

- [Wurzburg, Germany, 2011](#)
The 11th annual conference and annual general meeting was held October 10-16, 2011, hosted by the University of Wurzburg.
- [Zadar, Croatia, 2010](#)
The tenth annual conference and annual general meeting was held November 8-14, 2010, hosted by the University of Zadar.
- [Ann Arbor, 2009](#)
The ninth annual conference and annual general meeting was held November 11-15, 2009, hosted by the University of Michigan, Ann Arbor.
- [London, 2008](#)
The eighth annual conference and annual general meeting was held on November 6-8, 2008, hosted by King's College London.
- [College Park, Maryland, 2007](#)
The seventh annual conference and annual general meeting was held on November 1-2, 2007, hosted by the University of Maryland in College Park, Maryland, USA.
- [Victoria, British Columbia, 2006](#)
The sixth annual conference and annual general meeting was held on October 27-28, 2006, in Victoria, Canada, and hosted by the University of Victoria.
- [Sofia, Bulgaria, 2005](#)
The fifth annual conference and annual general meeting was held on October 28-29, 2005, hosted by the Bulgarian Academy of Sciences in Sofia, Bulgaria.
- [Baltimore, Maryland, 2004](#)
The fourth annual conference and annual general meeting was held on October 21-22, 2005, hosted by Johns Hopkins University in Baltimore, USA.
- [Nancy, France, 2003](#)
The third annual conference and annual general meeting was held on November 7-8, 2003, hosted by ATILF in Nancy, France.
- [Chicago, Illinois, 2002](#)
The second annual conference and annual general meeting was held on October 11-12 2002, at the Newberry Library in Chicago, USA.
- [Pisa, Italy, 2001](#)
The first annual conference and annual general meeting was held on November 16-17, 2001 in Pisa, Italy.

TEI2022 Conference Programme

The overall schedule

As noted in the joining instructions above, the full programme of the conference is available at: <https://www.conftool.pro/tei2022/sessions.php>.

The general structure of the conference is:

Times / Days	Monday 12 September 2022	Tuesday 13 September 2022	Wednesday 14 September 2022	Thursday 15 September 2022	Friday 16 September 2022
09:00 - 09:30	Refreshments in King's Hall				
09:30 - 11:00	Workshops	Workshops	Session 1	Session 4	Session 7
11:00 - 11:30	Break in King's Hall				
11:30 - 13:00	Workshops	Workshops	Session 2	Session 5	Session 8
13:00 - 14:30	Buffet Lunch in King's Hall				
14:30 - 16:00	Workshops	SIG Meetings	Session 3	Session 6	Closing Keynote in ARMB.2.98
16:00 - 16:30	Break in King's Hall				Reception in King's Hall
16:30 - 18:00	Workshops	SIG Meetings	Poster Session in King's Hall	TEI-C AGM in ARMB.2.98	
18:15 - 19:30		Opening Keynote in ARMB.2.98			
19:30 - 21:00		Reception in King's Hall			

Keynotes Lectures

Opening Keynote: Constance Crompton, "Situated, Partial, Common, Shared: TEI Data as Capta"

Time: Tuesday, 13/Sept/2022: 6:15pm - 7:30pm

Session Chair: James Cummings, Newcastle University

Location: ARMB: 2.98

Armstrong Building: Lecture Room 2.98.

Starting with: Welcome To Newcastle University, Professor Jennifer Richards, Director of the Newcastle University Humanities Research Institute.

Situated, Partial, Common, Shared: TEI Data as Capta

C. Crompton

University of Ottawa, Canada

Abstract:

It has been a decade since Johanna Drucker reminded us that all data are capta in the TEI-encoded pages of Digital Humanities Quarterly. In some ways this may appear to be self-evident in the context of the TEI: for many TEI users, their primary encoded material is text, and the TEI tags are a textual intervention in the sea of primary text – the resulting markup is not data, as in something objectively observed, but rather capta, as in something situated, partial, and contextually freighted (as indeed is all data. All data is capta). That said, Drucker warns her readers against self-evident claims.

Drawing on Drucker's arguments, this keynote explores the tension in several of the TEI's models, and the challenges that arise from our need to have fixed start and end points, bounding boxes, interps, certainty, events, traits (the list goes on!) in order to do our analytical work. Drawing on a number of projects, I argue for the value of our shared markup language and the value it offers us through its data-like behaviour, even as it foregrounds how clearly how much TEI data, and indeed, all data, are capta.

Closing Keynote: Emmanuel Ngue Um, 'Tone as "Noiseless Data": Insight from Niger-Congo Tone Languages'

Time: Friday, 16/Sept/2022: 2:30pm - 4:00pm

Session Chair: Martina Scholger, University of Graz

Location: ARMB: 2.98

Tone as "Noiseless Data": Insight from Niger-Congo Tone Languages

E. Ngue Um

University of Yaoundé 1 & University of Bertoua (Cameroon), Cameroon

With Closing Remarks, Dr James Cummings, Local TEI2022 Conference Organiser

Abstract:

Text processing assumes two layers of textual data: a "noisy" layer and a "noiseless" layer. The "noisy" layer is generally considered unsuitable for analysis and is eliminated at the pre-processing stage. In current Natural Language Processing (NLP) technologies like text generation in machine translation, the representation of tones as diacritical symbols in the orthography of Niger-Congo languages leads to these symbols being pre-processed as "noisy" data. As an illustration, none of the 15 Niger-Congo tone languages modules available on Google Translate delivers in a systematic and consistent manner, text data that contains linguistic information encoded through tone melody.

According to Yip (2002), 60 to 70% of the world's languages are tonal. In Niger-Congo languages, Tone is a pitch melody which is associated with the articulation of tone bearing units (e.g., vowels, syllables). Such is the case with Basaa, a Bantu language spoken in Cameroon. In the description of the Basaa language, tones are analyzed as structural units both in the phonology and in the grammar. At the level of the phonology, two seemingly homophonous and/or homographic words, for example **hím** "to jump" and **h̃m** "to mumble", have different meanings on the basis of the difference of the pitch of their respective syllables. At the syntactic level, variation in the pitch of a syllable may signal new grammatical information like tense, aspect, clause type, prosody, as in examples 1 and 2 below.

1a.	mɛ̀	hím [L H]	1b.	mɛ́	h̃m [HHL]
	3SG	jump.PST			1SG.PST CONJ.jump.
		"I jumped"			"let me jump!"
2a.	mɛ̀	h̃m [L LH]	2b.	mɛ́	h̃m* [H H↑L]
	3SG	jump.PST			1SG.PST CONJ.jump.
		"I mumbled"			"let me mumble!"

where L, H, HL, LH H↑L stand for "low", "high", "high-low", "low-high", and "high-upstep low" pitches, respectively. The asterisk in 2b signals a non phonetic representation of the tone melody on the verb **h̃m** "to mumble", because the graphical layout of HL as a circumflex does not provide for spacing and insertion of the upstep symbol [↑]. A workaround representation would be to double the vowel which is the tone bearing unit, as in 2c

- 2c. **mɛ́ hɪ̃im*** [H H↑L]
 1SG.PST CONJ.jump.
 “let me mumble!”

In the standard orthography of some Niger-Congo languages, “accidentals” such as upstep and downstep are represented by a mid-tone character.

Graphical representation of tone as diacritical symbols is “noisy” with regard to at least three aspects :

- a. Fuzziness of the information being encoded, like when acute accent representing high tone encodes both lexical and grammatical information, as in 1a (hím [H]) and 1b (mɛ́[H]), respectively.
- b. Tokenization and parsing of tone characters in combining diacritics, as in 1b (hím [HL]), 2a (hĩm [LH]), and 2b (h̃im* [H↑L]).
- c. Redundancy of tone marking and melodic “accidentals” which affect tone melody, like upstep and downstep. Redundancy concerning tone marking lies in the fact that only a small subset of a specific tone language’s lexicon would require tone marking, if tone orthography were informed by quantitative study of the tone. Regarding melodic “accidentals” like the upward-arrow which is placed before the second [i] vowel bearing a low tone in 2c, it indicates that the pitch of the low tone on the [i] vowel is raised by some degree, as a result of the presence of a preceding high tone (i.e, the tone placed on the first [i] symbol in 2c). This implies that upstep here is a purely phonetic phenomenon which signals modulation of the tone register but does not encode grammatical information per se. The only justification for marking upstep in 2c would be to guide the reader's pronunciation. Therefore, such a mark can justifiably qualify as “noisy”.

The Text Encoding Initiative (TEI) is a framework which can be used to circumvent the “noisiness” brought about by diacritical tone symbols in the processing of text data of Niger-Congo languages.

In novel work, I propose a markup scheme for tone that encompasses:

- a. The markup of tone units within an <m> (morpheme) element; this aims to capture the functional properties of tone units, just like segmental morphemes.
- b. The markup of tonal characters (diacritical symbols) within a <g> (glyph) element and the representation of the pitch by hexadecimal data representing the Unicode character code for that pitch; this aims to capture tone marks as autonomous symbols, in contrast with their combining layout when represented as diacritics.
- c. The markup of downstep and upstep within an <accid> (accidental) element mirroring musical accidentals such as “sharp” and “flat”; this aims to capture strictly melodic properties of tone on a separate annotation tier.

The objectives of tone encoding within the TEI framework are threefold:

- a. To harness quantitative research on tone in Niger-Congo languages.
- b. To leverage “clean” language data of Niger-Congo languages that can be used more efficiently in machine learning tasks for tone generation in textual data.
- c. To gain better insights into the orthography of tone in Niger-Congo languages.

In this paper, I will show how this novel perspective to the annotation of tone can be applied productively, using a corpus of language data stemming from 120 Niger-Congo languages.

Workshops

Workshop 1: From a collection of documents to a published edition : how to use an end-to-end publication pipeline [Full Day]

Time: Monday, 12/Sept/2022: 9:30am - 6:00pm

Location: ARMB: 3.38

Keywords: digital edition, historical manuscripts, encoding pipeline, publication workflow

In 2021, during the last edition of the TEI Conference “Next Gen TEI”, I took part in a session where I presented a project I had been working on for a year and a half. This project, both relying massively on the Text Encoding Initiative and benefiting its community, focusses on the creation of a pipeline for the publication of digital scholarly editions. This pipeline, which was still a work in progress at the time of the 2021 Conference, but is now complete, aims at providing open-source, free, easy-to-use and interoperable tools; its goal is to support the editorial process from the digitization of a collection of documents to its publication in a machine-readable standard.

In the following, I will succinctly describe the six steps that compose this pipeline, and then move to the way I intend to conduct the workshop based on them.

Firstly, the collection of images that composes the corpus has to be stored and curated somewhere online both to keep them available for researchers and for publication. For this task, we rely on [IIIF](#), to ensure sustainability and interoperability.

The three following steps, segmentation, transcription and post-OCR correction, are performed with [eScriptorium](#), an open-source transcription application. It offers various features: uploading images, production of ground truths, manual or automatic segmentation and transcription, using custom models, training segmentation and transcription models, to name a few. Finally, if there are any remaining errors in the transcription (in case of an automatic transcription), it is possible to either correct them manually in eScriptorium or export the files and correct them with the help of specifically designed scripts.

Once the transcription is fully done, we encode it in TEI XML. For this step, we provide various solutions, depending on the transcription file format (Page XML, XML ALTO, Text) chosen when exporting the transcription from eScriptorium. We also propose documented scripts that help automatize and speed up this process.

Encoded files are then published online with the help of TEI Publisher, an application designed for generating custom editions for corpora encoded in TEI XML. We have developed and launched a dedicated application for digital scholarly editions ([DiScholEd](#)) on this basis. It is available online together with a thorough [documentation](#), and is conceived as an open application: new corpora can always be added to it, and we welcome new collaborations.

The goal of our workshop is to demonstrate how a corpus could be processed for publication with TEI Publisher. The workshop participants will learn to experiment with a ready-to-use solution that provides an easy and quick publication of a corpus. They will also get tips and shortcuts to help speed up the creation of a digital edition. Moreover, by the end of the session, this workshop will provide the participants with a visualization of their respective corpus, with side by side transformed text and original image; all of which then showing what can be achieved while working with TEI in the context of an end-to-end publication pipeline. The program for this workshop is the following: firstly, it will start with a presentation of the pipeline, its objectives and how it works. Then, the time we have will be divided into several slots corresponding to every step of the pipeline. Each slot will start with a quick presentation of what is expected of the participants and what tools they will need to use. Next, they will be allotted some time to work with their data and to process them for publication. At the end of the day, a 30mn feedback session will make it possible for each participant as well as for the workshop organizers to assess the benefits of the session and envision further possible collaborations.

Considering the number of steps in this pipeline and the time required for each of these steps, a full day is necessary for this workshop. The number of participants should be 10-15 maximum, in order for the two workshop conveners to be able to provide the necessary technical support for the hands-on parts of the workshop.

In order for the participants to be able to work correctly on the pipeline, they will need a laptop as well as the following tools: a command line interface for the execution of the scripts, an XML editor ([Oxygen](#) is the best choice) and a way to work with TEI Publisher. The latter can be launched with a local eXist-db installation, or with docker (see [Documentation](#)). An eScriptorium account will be provided to each participant. They will also have to bring their own material (textual sources preferably; images and transcription (in TXT format)) to work on (about 3 to 5 pages).

GitHub repository of the pipeline:

<https://github.com/DiScholEd/pipeline-digital-scholarly-editions>

Workshop leader(s)

Floriane Chiffolleau:

After a master's degree in late modern history and in "Technologies numériques appliquées à l'histoire" at the Ecole nationale des Chartes, Floriane Chiffolleau worked as a research and development engineer at Inria for a year and a half. She then started a PhD in digital humanities in October 2021 under the direction of Anne Baillet at Le Mans Université (3L.AM) and Laurent Romary at Inria (ALMAnaCH). Her research focuses on text recognition and TEI encoding.

Email: floriane.chiffolleau@inria.fr

Hugo Scheithauer:

Research and Development Engineer in the Inria ALMAnaCH team, Hugo Scheithauer holds a master's degree in art history and in “Technologies numériques appliquées à l'histoire” at the École nationale des chartes. He works on the automatic segmentation of sale catalogues for the DataCatalogue project, jointly led by Inria, the National Library of France (BnF) and the National Institute for Art History (INHA).

Email: hugo.scheithauer@inria.fr

Workshop 2: Creating Digital Editions with FairCopy [Half Day, Afternoon]

Time: Monday, 12/Sept/2022: 2:30pm - 6:00pm

Location: ARMB: 1.04

Keywords: Digital Humanities Critical Editions Tools IIF

Abstract

In this half day workshop, participants will learn how to use FairCopy to transform historical texts into online digital editions. Using crowdsourced transcriptions as a starting point, we will add semantic structure and mark names of people, places, and events. We will then publish our digital editions using Hugo.

Introducing FairCopy

The TEI Guidelines have been used by hundreds of scholarly projects and are an essential tool for researching, preserving, and disseminating cultural heritage world-wide. And yet, despite its mission to provide a common vocabulary for describing texts, TEI faces problems of adoption and use in the wider scholarly community. While the basics of TEI XML encoding are simple enough, true fluency in TEI requires institutional support and commitment in the form of training, technical staff, IT infrastructure, and the time and commitment of the individual scholar.

Even within institutions that have these resources, projects often adopt a simpler interface for domain experts to interact with. This interface then translates the scholar's work into TEI behind the scenes. This is sometimes accomplished technologically, sometimes through a tiered system of labor, or both. These interfaces are more often than not specialized to the needs of the projects which develop them. This current state of affairs leads to a structural problem of access which further limits whose texts can be digitized and preserved.

FairCopy addresses this problem of access by providing a simple editing environment in which anyone can produce valid TEI documents. FairCopy doesn't hide the complexity of TEI, but rather makes it available for users to explore at their own pace. Users are quickly comfortable with its interface and able to focus on the text, not XML syntax.

FairCopy has support for most of the 500+ elements in TEI and allows users to customize a schema for their particular project. Scholars can seamlessly import and export TEI-XML documents. Additionally, scholars can bring in IIF images of primary resources and link them to their transcriptions.

The Workshop

In this half day workshop, participants will learn how to use FairCopy to transform historical texts into online digital editions encoded using TEI. Using crowdsourced transcriptions as a starting point, we will add semantic structure and mark names of people, places, and events. We will then publish our digital editions using Hugo.

In the first part of the workshop, we will begin with a demonstration of FairCopy. We will then select texts to work on based on participants interests. Participants are encouraged to bring their own texts. Finally, we will break into small groups.

In the second part, each group will work on encoding a text using FairCopy. Participants will work collaboratively to choose elements and attributes that best suit their selected texts. The presenter will float between groups answering questions.

In the third part, we will export our texts into a premade Hugo template that can display both the original IIF page images and the TEI encoded texts.

Participants in this workshop will need to bring a Mac, Windows, or Linux laptop on which they can install FairCopy for free. No web design or XML skills are required.

Participants in this workshop will learn how to use FairCopy to create a digital edition. They will also learn about using TEI semantics to structure and mark texts. They will also gain familiarity with using IIF Manifests to interoperate between library collections and digital editions.

Speaker Bio

Nick Laiacona is a partner at [Performant Software Solutions LLC](#). Performant serves clients in the Digital Humanities throughout North America and Europe. Laiacona has developed tools for critical digital editions including: Juxta, Digital Mappa, TextLab, and now FairCopy. Laiacona has helped produce a number of critical editions, including “Secrets of Craft and Nature in Renaissance France” and the “Melville Electronic Library.”

Workshop 3: A short introduction to Schematron [Half Day, Afternoon]

Time: Monday, 12/Sept/2022: 2:30pm - 6:00pm

Location: ARMB: 1.06

Keywords: Schematron, Validation, Quality Assurance

Schematron is a rule based validation language for structured documents. It was designed by Rick Jelliffe in 1999 and standardized as ISO/IEC 19757-3 in 2006. It lets you evaluate assertion test for selected parts of a document. Schematron uses XPath both as the language to select the portion of a document and as the language of the assertion tests. This use of XPath gives Schematron the flexibility to validate arbitrary relationships and dependencies of information items in a document.

What also sets Schematron apart from other languages is that it encourages the use of natural language descriptions targeted to human readers. This way validation can be more than just a binary distinction (document valid/invalid) but also support authors of in-progress documents with quick feedback on erroneous or unwanted document structure and content.

The flexibility and (relative) simplicity of Schematron makes it an invaluable tool for XML-based text encoding projects. The range of supported tasks reaches from "hard" validation to enforce constraints on documents to "soft" validation to report potential problems such as Unicode characters from Unicode Private Use Areas to interactive error correction with Schematron extensions like Schematron QuickFix.

This half-day workshop will introduce the participants to the principal idea of Schematron and practice its application to XML-based text encoding projects. Together we will explore patterns, rules, and assertions as the basic Schematron concepts and touch phases, variables, and abstract patterns as more advanced features of Schematron validation.

From the participants the workshop requires a general understanding of XML document editing and knowledge of XPath. The material requirements are a projector and laptops to follow through with the examples given in the workshop. Any operating system with a recent Java Runtime is sufficient

Participants are recommended to bring their own device.

Workshop Leader:

David Maus is head of research of development at the State and University Library Hamburg. He acts as liaison to digital humanities research at the University of Hamburg and other higher education institutions. He is currently deeply involved as information architect and XML programmer in Dehmel Digital, a digital scholarly edition that uses machine-learning technologies to provide access to the correspondence of Richard and Ida Dehmel, a famous artist couple around 1900. He is the author of SchXslt, a modern implementation of the Schematron validation language for structured documents and serves on the program committee of the MarkupUK conference.

Workshop 4: Building TEI-powered websites with static site technology. A hands on exploration of the publishing toolkit of the Scholarly Editing Journal [Half Day, Morning]

Time: Tuesday, 13/Sept/2022: 9:30am - 1:00pm

Location: ARMB: 3.38

Keywords: Digital publishing, TEI processing, static sites, programming

Raffaele Vigiante, Maryland Institute for Technology in the Humanities, University of Maryland

This half-day (approximately 3 hours) workshop will introduce TEI publishing with static site generators and front-end technologies, namely the static site generator Gatsby and React JS. It will introduce the attendees to the publishing strategies and tool sets developed for the reboot of the online Scholarly Editing journal (<https://scholarlyediting.org/>), which publishes, among essay-like content, TEI-based small scale editions. This workshop is aimed at attendees who already have some experience with programming (including XSLT) and the command line; however, all are welcome and will be supported as much as possible throughout the workshop.

The publishing tools presented in this workshop were developed for the reboot of the Scholarly Editing journal, which published its newest issue, volume 39, in April 2022. The previous site, built with Apache Cocoon, was converted into a static site and made accessible as an archive (<https://scholarlyediting.org/se.index.issues.html>). The new website and journal issues are built using Gatsby, a static site generator that relies on React JS for building user interfaces. The journal's editors chose to adopt a static site generator because, once built, static sites do not need maintenance and can be easily moved and archived. This requires less infrastructure to publish and keep the site online on the web, which is desirable both for keeping operational costs of the journal low and to ensure its longevity. XML technologies can be and are used to generate static sites; the TEI Guidelines are a notable example. Regardless of how the static site is built, the result has minimal infrastructure requirements. A server is always needed to publish something on the web, but its role is limited to sending files over to the client, essentially just supporting HTTP GET operations. This is cheap and it makes it possible to rely on affordable web hosting, or take advantage of free services, or even use a home server.

During the workshop, participants will create a Gatsby website starting from a provided template that includes the TEI rendering tools `gatsby-transformer-ceteicean` and `gatsby-theme-ceteicean`. These tools re-implement principles pioneered by CETEIcean, which relies on the browser's DOM processing and HTML5 Custom Elements to publish TEI documents as a component pluggable into any HTML structure (Cayless and Viglianti 2018). Example TEI documents to integrate into the website will be provided, but attendees are encouraged to bring their own.

After an introduction on static sites, the motivations for using them, and an open discussion, the workshop will introduce:

- How to set up Gatsby and the CETEIcean plugins
- How to use built-in behaviors
- Customization via CSS (and CSS-in-JS)
- Defining custom behaviors as React components (we will cover the fundamentals of React components in this workshop)
- Applying optional transformations to TEI documents before and after ingestion into Gatsby (as we will see, transformations must be run by NodeJS and therefore written in JavaScript; however, XSLT scripts can also be applied via SaxonJS).

If time allows, we will conclude with open discussion and collaborative experimentation.

Participants must bring their own laptop and be able to install (free) software on it. Internet access will be required. The tutor will require a projector.

References:

Cayless, Hugh, and Raffaele Viglianti. "CETELcean: TEI in the Browser." Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21 (2018). <https://doi.org/10.4242/BalisageVol21.Cayless01>.

Biography:

Dr. Raffaele (Raff) Viglianti is a Senior Research Software Developer at the Maryland Institute for Technology in the Humanities, University of Maryland. His research is grounded in digital humanities and textual scholarship, where "text" includes musical notation. He researches new and efficient practices to model and publish textual sources as innovative and sustainable digital scholarly resources. Dr. Viglianti is currently an elected member of the Text Encoding Initiative technical council and the Technical Editor of the *Scholarly Editing* journal.

Workshop 5: Introduction to XProc [Half Day, Morning]

Time: Tuesday, 13/Sept/2022: 9:30am - 1:00pm

Location: ARMB: 3.41

Keywords: XProc, Automation, Pipeline

XProc is an XML based programming language for processing documents in pipelines. Version 1.0 of the language was published as W3C Recommendation in 2010. The final version of the next version, XProc 3.0, is expected to be published as community report in late 2022.

This half-day workshop will teach the participants the basic concepts of an XProc processing pipeline (pipelines, steps, ports) and practice their application in a series of exercises. The overall goal of the workshop is to enable participants to write pipelines that chain common markup manipulation tasks such as loading, transforming, validating that can be used as building blocks for more elaborate steps or as one-off scripts in data maintenance.

From the participants the workshop requires a general understanding of XML document editing and basic knowledge of XPath. The material requirements are a projector and laptops to follow through with the examples given in the workshop. Any operating system with a recent Java Runtime is sufficient.

Workshop Leader:

David Maus is head of research of development at the State and University Library Hamburg. He acts as liaison to digital humanities research at the University of Hamburg and other higher education institutions. He is currently deeply involved as information architect and XML programmer in Dehmel Digital, a digital scholarly edition that uses machine-learning technologies to provide access to the correspondence of Richard and Ida Dehmel, a famous artist couple around 1900. He is the author of SchXslt, a modern implementation of the

Schematron validation language for structured documents and serves on the program committee of the MarkupUK conference.

Workshop 6: Engaging TEI Editors Through LEAF-Writer [Half Day, Morning]

Time: Tuesday, 13/Sept/2022: 9:30am - 1:00pm

Location: ARMB: 1.06

Keywords: TEI-XML, web-based editor, RDF, named entity recognition

Engaging TEI Editors Through LEAF-Writer

Susan Brown (University of Guelph), Diane Jakacki (Bucknell University), James Cummings (Newcastle University), Mihaela Ilovan (University of Alberta), Luciano Frizzera (Canadian Writing Research Collaboratory), Rachel Milio (Bucknell University), Carolyn Black (LAB Cooperative)

The digital humanities community has worked for decades to establish best practices in the production and dissemination of richly annotated electronic texts, enabling scholars to analyze, share, and collaborate on literary cultural heritage materials. And yet, training in text encoding and access to expensive proprietary software is beyond the reach of many scholars and students. Therefore, a gap remains between those with encoding expertise and those without, and the opportunity to participate in transformative modes of transmission and open collaboration eludes many in our profession and in our classrooms.

More than ever, we need to find ways to bridge that gap and develop open-source, open-access tools that support all researchers and editors as they work on electronic texts, using accepted standards without having to take on additional technical labor and financial commitment in order to participate.

LEAF-Writer¹ offers one solution to this problem: an open-source, open-access Extensible Markup Language (XML) editor that runs in a web browser and offers scholars and students a rich textual editing experience without the need to download, install, and configure proprietary software, pay ongoing subscription fees, or learn complex coding languages. This user-friendly git-backed editing environment incorporates Text Encoding Initiative (TEI) and Resource Description Framework (RDF) standards, meaning that texts edited in LEAF-Writer are interoperable with other texts produced by the scholarly editing community and with other materials produced for the Semantic Web. LEAF-Writer is particularly valuable for pedagogical purposes, allowing instructors to teach students best practices for encoding texts without also having to teach students how to code in XML directly. LEAF-Writer is designed to help bridge the gap by providing access to all who want to engage in new and important forms of textual production, analysis, and discovery. LEAF-Writer draws on TEI All as well as other

¹ LEAF-Writer is a core component of the Linked Editorial Academic Framework virtual research environment that is available as a standalone module that can be run using a range of backend frameworks or embedded in larger environments. The open instance at <https://leaf-writer.stage.lincsproject.ca> allows users to work from PCs, or to use GitHub or GitLab for storage and collaboration.

TEI-C-supplied schemas, can use project-specific customized schemas, and offers continuous validation against supported and declared schemas. LEAF-Writer allows users to access and synchronize their documents in GitHub and GitLab, as well as to upload and save documents from their desktop.

In this half-day hands-on workshop, participants will learn how to make the most of LEAF-Writer's extensive capabilities. From their own laptops, they will learn how to choose among TEI customizations to best support their work in diplomatic and/or semantic markup, add inline scholarly notes and glosses, and create annotations - tagging named entities and associating them with recognized authorities like VIAF, Wikidata, and Getty - that do double duty as in-text identifiers and potential contributions to the Semantic Web. Workshop facilitators will model how LEAF-Writer can be employed in a variety of group collaborations as well as for classroom settings. Using supplied texts, templates, and their own materials, participants will be able to experiment with texts in a number of genres, including verse, drama, prose, and documentary archival materials, as well as texts in multiple languages. By the end of the workshop, participants will possess the skills to:

- Fully encode a document in LEAF-Writer, choosing amongst XML, RDF, or XML-RDF modes
- Leverage LEAF-Writer's built-in named entity recognition and vetting engine (NERVE) to more efficiently capture and tag with unambiguous identifiers people, places, and organizations within their texts
- Import texts into LEAF-Writer from the web or their own laptop
- Save their work to file sharing / versioning repositories like GitHub or GitLab to enable collaboration with peers or colleagues

Workshop participants will come away with sample texts, documentation and training materials, and models for how LEAF-Writer can be integrated into editorial production pipelines.

Requirements: participants will need a web-enabled laptop computer, and have the Chrome or Firefox browsers installed. Model texts in a variety of genres will be available for experimentation. Participants are welcome to bring their own XML texts to work on (schema declaration must point to an .rng version of one of the TEI consortium-supported schema customizations).

LEAF-Writer's ongoing development and access are made possible through generous funding from the Canada Foundation for Innovation, the Mellon Foundation, CANARIE, the Canada Research Chairs program, and the Social Sciences and Humanities Research Council.

Bios:

Susan Brown is Canada Research Chair in Collaborative Digital Scholarship and Professor of English at the University of Guelph. Her work focuses on intersectional feminist literary history, semantic technologies, and critical infrastructure studies. She directs the Orlando

Project, the Canadian Writing Research Collaboratory, and the Linked Infrastructure for Networked Cultural Scholarship. She is President (2022-23) of the Alliance of Digital Humanities Organizations.

Diane Jakacki is Digital Scholarship Coordinator and Associated Faculty in Comparative & Digital Humanities at Bucknell University. She is PI of the Mellon-funded LAB Cooperative project, co-PI (with James Cummings) of the Evolving Hands project, and member of the LEAF executive team. She is chair of the TEI Board of Directors and Chair-Elect of the Alliance for Digital Humanities Organizations.

James Cummings is the Senior Lecturer in Late Medieval Literature and Digital Humanities for the School of English Literature, Language, and Linguistics at Newcastle University. He is a member of the overall LEAF project, the TEI Board of Directors, and for some reason volunteered to be lead local organiser for the TEI2022 conference.

Mihaela Ilovan is the Assistant Director of the Canadian Writing Research Collaboratory. A librarian and digital humanist by formation, she is the technical project manager for the development of the LEAF Virtual Research Environment and the coordinator for LEAF data ingestion and mapping.

Luciano Frizzera is an Interface Designer and Web Developer at the Canadian Writing Research Collaboratory. He is currently a PhD candidate in Communication Studies at Concordia University, researching the processes of algorithmic subjectivation and mediation on digital platforms.

Carolyn Black is Associate Researcher for the LAB Cooperative project, and member of the LEAF project team.

SIG Meetings

Tuesday, 13/Sept/2022

2:30pm - 4:00pm

- SIG 1: Manuscripts
 - Location: ARMB: 3.38

- SIG 2: Ontologies
 - Location: ARMB: 1.06

- SIG 3: Linguistics
 - Location: ARMB: 3.41

4:30pm - 6:00pm

- SIG 4: Correspondence
 - Location: ARMB: 3.38

- SIG 5: Newspapers and Periodicals
 - Location: ARMB: 1.06

Conference Sessions – Wednesday 14 September 2022

Session 1A – Short Papers – 09:30 - 11:00

Session 1A: Short-Papers

Location: ARMB: 2.98

Chair: Martin Holmes, University of Victoria

Short Paper: Standoff-Tools. Generic services for building automatic annotation pipelines around existing tools for plain text analysis

C. Lück

Universität Münster, Germany

Keywords: text mining, stand-off annotations, models of text, generic services

Standoff-Tools: Generic services for building annotation pipelines around existing tools for plain text analysis

Christian Lück

TEI XML excels at encoding text. But when it comes to machine-based analysis of a corpus as data, XML is no good platform. NLP, NER, topic modeling, text-reuse detection, etc. work on plain text; they get complicated and slow if they traverse a tree structure. While extracting plain text from XML is simple, feeding the result back into the XML source document is tricky. However, having the analysis' result in there is desired: It can be related to the internal markup, e. g. for overviews of names per chapter. In my short paper, I will introduce standoff tools², a suite of generic tools for building annotation pipelines around any plain text tool, which provides character offsets like Spacy or ANTLR parsers. In addition, standoff tools can internalize stand-off markup produced manually with CATMA, GNU Emacs standoff-mode, etc. so that the result is syntactically correct XML (cf. Cayless 2019).

Standoff tools commit to clear design principles: Unlike competitors (Andorfer et al. 2021, Meyer 2022) they do not take any NLP tool as a fixed constant and do not import any language model (Schopper 2021, Banski et al. 2016). Unlike integrated platforms, e. g. TXM or OCRE, they are committed to the Unix/GNU philosophy and microservices: They just serve the tasks of extracting plain text for analysis with other tools and–most important–to re-integrate the result of the analysis back into the XML source document. They are schema-agnostic, i. e. TEI isn't hard-wired into the code. The core library is not preassigned to XML at all but offers abstractions for any tree-structured markup language. It does not apply X-technologies but XML parsing–in a non-standard way with character offsets of all nodes–and then it processes integers.

² <https://github.com/lueck/standoff-tools>

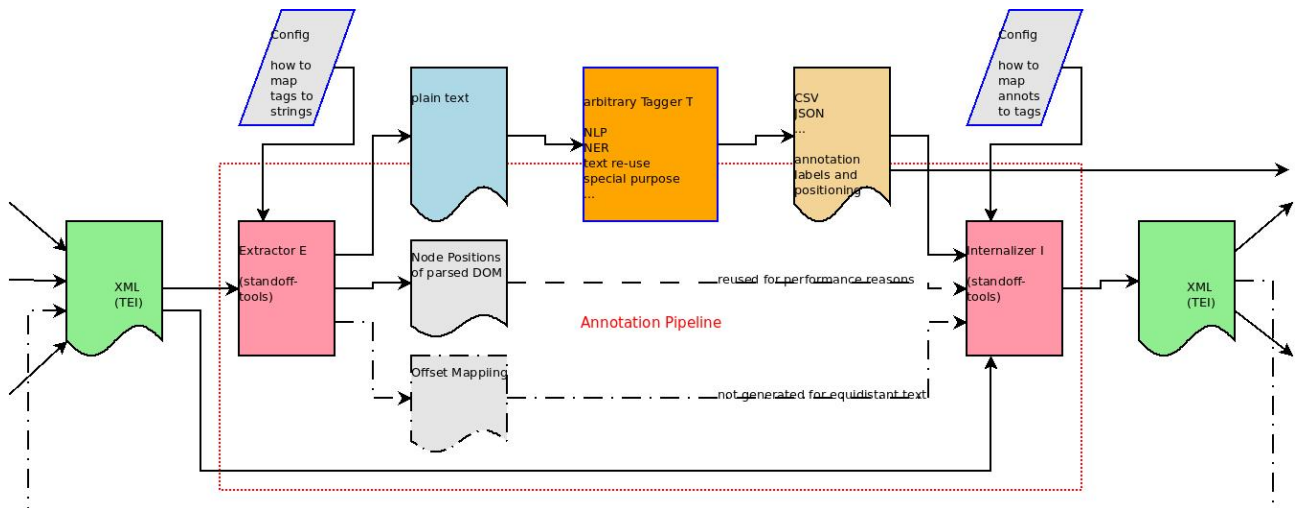


Fig.1. Annotation pipeline for TEI XML build with standoff-tools

In order to build an annotation pipeline, standoff tools provide pairs of two concerted programs, the extractor E and the internalizer I. (Fig. 1) The simplest pair is E and I for a special flavor of plain text, I term *equidistant plain text*: The XML tags and PIs are replaced by strings of equal length consisting of e. g. white space. Hence, all non-white-space characters in the plain text have the same character offsets as their correspondents in the XML document. If we feed this equidistant plain text to an arbitrary tagger T designed for plain text, then I can directly use the provided character references to enrich the XML document. When we want entity and character refs to be expanded or tag-interleaved words to be joined, or when we need tags like `<tei:p>` to be replaced with double newlines, a more refined pair of E and I is required. It is impossible to concert this pair through equidistant plain text. Instead, E now has to keep track of all the replacements that go into the plain text. Therefore, it generates a *mapping of offsets* that maps each character in plain text to positions in XML (a vector of integers). Using this mapping, I can unambiguously translate the offsets produced by the tagger back to XML offsets.

The internalizer I produces syntactically correct XML by splitting the ranges to be annotated. It does not produce embedded stand off annotations (Banski et al. 2016), which would be a trivial algorithm, but inline annotations. They are split if there would be overlapping edges otherwise. Splitting and inserting open/close tags is based on integer calculations. There is no need for slow backtracking algorithms. It performs very well even on large amounts of annotations overlapping each other and the internal markup. While splitting, I produces data for keeping track of elements that originate from the same annotation and can hence be configured to re-aggregate them through `@prev` and `@next` attributes (cf. TEI 2022, [chap. 16.7](#)). By reason of its performance and the inline annotations, the internalizer alone can also serve as a component for visualizing OA-based stand off annotations.

References

- Andorfer, Peter and Schlögl, Matthias (2021): acdh-spacytei. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt “Kompetenznetzwerk Digitale Edition”. Handle: hdl.handle.net/11471/562.50.2. Accessed on 2022-08-03.
- Banski, Piotr et al. (2016): Wake up, standOff! In: HAL CCSD 2016, oai:HAL:hal-01374102v1.

- Cayless, Hugh (2019): Implementing TEI Standoff Annotation in the browser. In: Proceedings of Balisage: The Markup Conference 2019, no. 23, doi: [10.4242/BalisageVol23.Cayless01](https://doi.org/10.4242/BalisageVol23.Cayless01) Accessed on 2022-08-03.
- Meier, Wolfgang (2022): Names sell. Named Entity Recognition in TEI Publisher, <https://e-editiones.org/blog/> Accessed on 2022-08-03.
 - Schopper, Daniel (2021): xsl-tokenizer. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt “Kompetenznetzwerk Digitale Edition”. Handle: hdl.handle.net/11471/562.50.216. Accessed on 2022-08-03.
 - The TEI Consortium (2022): TEI P5. Guidelines for Electronic Text Encoding and Interchange. Version 4.4.0.

Short Paper: TEI Automatic Enriched List of Names (TAELN): An XQuery-based Open Source Solution for the Automatic Creation of Indexes from TEI and RDF Data

G. Fernandez Riva

Universität Heidelberg, Germany

Keywords: TEI, indexes, XQuery

TEI Automatic Enriched List of Names (TAELN): An XQuery-based Open Source Solution for the Automatic Creation of Indexes from TEI and RDF Data.

Gustavo Fernández Riva

Universität Heidelberg – SFB 933 “Materiale Textkulturen“

The annotation of names of persons, place or organizations is a common feature of TEI editions. One way of identifying the annotated individuals is through the use of IDs from authority records like Geonames, Wikidata or the GND.

In this paper I will introduce an open source tool written in XQuery that enables the creation of TEI indexes using a very flexible custom templating language. The TEI Automatic Enriched List of Names (TAELN) uses the ids according to one authority document to create a file for an index (model.listLike) with information from one or more RDF endpoints.

TAELN has been developed for the edition of texts from Albrecht Dürer and his family. People, places and works of art are identified with GND-numbers in the TEI edition. The indexes generated include some information from GND records, but mostly from duerer.online¹, a virtual research portal, created with [WissKI](https://www.wisski.de/)², which includes an RDF endpoint.

TAELN relies on XML-templates to indicate the information to retrieve from the different endpoints as well as how to structure the desired TEI output. They use a straight-forward but flexible and powerful syntax. A simple use case is depicted in the following example that retrieves the person name from the GND and the occupation from [WissKI](https://www.wisski.de/) (using the so-called »Pathbuilder syntax«).

```
<lb n="16" />desiderato <persName
ref="pers:gnd-11852786X">Alberto</persName>, testis Liber,
```

Figure 1: Excerpt of the source file in TEI. Proper names are encoded with `persName` and a reference that includes their number in the GND authority file.

```
<person>
  <persName origin="gnd">preferredNameForThePerson</persName>
  <occupation origin="wisski">ecrm:E21_Person ->
ecrm:P11i_participated_in -> wvz:WV7_Occupation ->
ecrm:P3_hab_note</occupation>
</person>
```

Figure 2: Excerpt of the configuration file, where the connections between source file, RDF-Endpoints and output file are described.

```
<person>
  <persName>Albrecht Dürer</persName>
  <occupation>Maler</occupation>
</person>
```

Figure 3: The result in the index TEI file.

Much more complex outputs can be achieved. TAELEN offers editions an out of the box solution to generate TEI indexes by gathering information from different endpoints and it only requires the creation of the corresponding template and a beginner knowledge of XQuery transformations. The tool is part of the heiEDITIONS³ framework for digital editions developed at the University of Heidelberg and the code is maintained in GitHub⁴. At the moment it is designed to use the GND numbers as identifiers and can only retrieve information from the GND and any WissKI system, but other projects could expand the code to include other authority files and RDF-endpoints.

Notes:

1 [duerer.online](https://sempub.ub.uni-heidelberg.de/duerer.online/): Virtuelles Forschungsnetzwerk Albrecht Dürer.

2 <https://sempub.ub.uni-heidelberg.de/duerer.online/>

3 <https://wiss-ki.eu/>

4 <https://heieditions.github.io/>

5 <https://github.com/GusRiva/taeln>

Short Paper: manuForma – A Web Tool for Cataloging Manuscript Data

M. de Molière

University of Munich, Germany

Keywords: manuscripts, codicology, paleography, XForms

manuForma – A Web Tool for Cataloging Manuscript Data

Maximilian de Molière

University of Munich, Germany; maximilian.moliere@lmu.de

The team of the ERC-funded project "MAJLIS – The Transformation of Jewish Literature in Arabic in the Islamic World" at the University of Munich needed a software solution capable of describing manuscripts in TEI-XML and that would be easy to learn for non-specialists. After about one year of development, manuForma now provides our manuscript catalogers with an accessible platform for entering their data. The elements (along with their attributes) are displayed as building blocks that users can pick from a list and rearrange them using an intuitive button interface. Another way manuForma tries to ease users into working with TEI is through comprehensive tooltips which explain the function of every element of the project and translates them into concepts familiar to scholars in our field. While manuForma does not spare our catalogers the need to learn the fundamentals of TEI, the restrictions the forms-based approach proffers enhance both TEI conformance and the uniformity of our catalog records. Moreover, our tool eliminates the need to install commercial XML editors on the computer of every project member tasked with describing manuscripts. Instead, our tool offers a web interface for the entire editorial process.

At the heart, manuForma uses XForms, which has been modified to allow adding, moving and deleting elements and attributes. A tightly knit schema file controls which elements and attributes can be added and in which situations to ensure conformance to the project's scholarly objectives. As an existDB application, manuForma integrates well with other apps that provide the front end to the manuscript catalog. TEI-XML records can be stored on and retrieved from GitHub, tying the efforts of the entire team together. Our tool offers user management to ensure that every team member gets credit for the records they collaborated on. The web solution is adaptable to other entities by writing a dedicated schema and a template file. Moreover, manuForma will be available under an Open Source license.

Session 1B – Long Papers – 09:30 - 11:00

Session 1B: Long Papers

Time: Wednesday, 14/Sept/2022: 9:30am - 11:00am

Chair: Syd Bauman, Northeastern University

Location: ARMB: 2.16

Long Paper: Texts All the Way Down: The Intertextual Networks Project

S. Connell, A. Clark

Northeastern University, United States of America

Title: Texts All the Way Down: The Intertextual Networks Project

Authors:

Sarah Connell, Northeastern University, sa.connell@northeastern.edu

Ash Clark, Northeastern University, as.clark@northeastern.edu

Keywords: intertextuality, bibliography, interface development, customization

Abstract:

In 2016, the Women Writers Project (WWP) began a research project on the multivalent ways that early women writers engaged with literate culture, at the center of which were systemic enhancements to a longstanding TEI corpus. Women Writers Online (WVO), collects approximately 450 works from the sixteenth to the nineteenth centuries, a watershed period in which women's participation in the authorship and consumption of texts expanded dramatically. With funding from the National Endowment for the Humanities, we used WVO's TEI encoding to jumpstart the creation of a standalone bibliography containing and linking to all the works referenced in WVO. This bibliography currently includes 3,431 book-level entries; 942 entries that are parts of larger works, such as individual essays or poems; and 126 simple bibliographic entries (e.g., books of the Bible). The bibliography identifies the genre of each work and the gender of the author, where known. We also expanded WVO's custom TEI markup to say more about "intertextual gestures"—or WVO authors' engagement with other works—which include not only named titles and quotations but also textual remix, adaptation, and parody. At time of writing, we have identified:

- 12,185 quotations, encoded in <quote>
- 5,570 direct title references, encoded in <title>
- 387 indirect title references, encoded in <rs> with @type of "title"
- 5,028 biblical references, encoded with the WWP's <regMe> element, wrapped in a <bibl>
- 2,035 other bibliographic references, encoded in <bibl>

Now, the WWP has published "Women Writers: Intertextual Networks"

(<https://wvp.northeastern.edu/intertextual-networks>), a web interface built on these two sources of rich TEI data: the bibliography and WVO's newly refined intertextual gestures. In this paper we will discuss the challenge of turning dense, textually-embedded data into an interface. Though the encoded texts themselves can stand alone as complete documents, we built Intertextual Networks with a focus on connective tissue, using faceting and linkages to invite curiosity about how authors and works are in conversation with each other. As the

numbers above suggest, this project attempts to enable investigations at scale, but we have also sought to draw out the local, even individual, ways that our writers engaged with other authors and texts. Thus, the interface includes visualizations that show patterns of usage (for example, the kinds of intertextual gestures employed by each author), but it also allows the reader to view the complete text of each gesture, reading through quotations, named titles, citations, and so on in full, with filtering and faceting to support exploration.

An important challenge for this project has been to build an interface that can address the multidirectional levels of textual imbrication at stake, allowing researchers to examine patterns among both referenced and referencing texts. This paper will share some key insights for TEI projects seeking to undertake similar markup expansion and interface development initiatives. We will discuss strategies for modeling, enabling discovery, and revealing complex layers of textual data and textuality among not only a primary corpus but also a related collection of texts.

Author biographies:

Ash Clark (e/em/eir) serves as the XML Applications Developer for the Northeastern University Women Writers Project and Digital Scholarship Group. Eir work focuses on inclusive and accessible markup, metadata, and web publications. E is currently working on Digital Humanities Quarterly's Biblio project, which seeks to map citations in the journal. E is also working on improved versions of Women Writers Online and the TEI Archiving, Publishing, and Access Service (TAPAS), a free publication hub for TEI documents.

Sarah Connell (she/her/hers) is the Associate Director of the Women Writers Project and the NULab for Texts, Maps, and Networks at Northeastern University. Her research focuses on text encoding and text analysis, medieval and early modern historiography, and pedagogies of digital scholarship. Her current projects include Making Room in History, a text encoding project on early modern narratives of national identity; Word Vectors for the Thoughtful Humanist, an NEH-funded seminar series on research and teaching with word embedding models; and "Representing Racial Identity in Early Women's Writing," a project examining discourses around race in the Women Writers Online collection.

Long Paper: Revising Sex and Gender in the TEI Guidelines

E. Beshero-Bondar¹, R. Viglianti², H. Bermúdez Sabel³, J. Jenstad⁴

1: Penn State Behrend, United States of America; 2: University of Maryland, United States of America; 3: University of Neuchâtel, Switzerland; 4: University of Victoria, Canada

Keywords: sex, gender, TEI Guidelines, document data, theory

Revising Sex and Gender in the TEI Guidelines

In Spring 2022, the co-authors collaborated in a TEI Technical Council subgroup to introduce a long-awaited <gender> element and attribute. In the process, we wrote new language for the *TEI Guidelines* on how to approach these concepts. As we submit this abstract, our proposed changes are under review by the Council for introduction in the next release of the *TEI Guidelines*, slated for October 2022. We wish to discuss this work with the TEI community to validate and address

- the history of the Guidelines' representation of these concepts,
- applications of the new encoding, and
- the extent to which the new specifications preserve backwards compatibility.

We must recognize as digital humanists and textual scholars that coding sex and gender as true “data” from texts significantly risks categorical determinism and normative cultural bias (Sedgwick 1990, 27+). Nevertheless, we believe that the TEI community is well prepared to encounter these risks with diligent study and expertise on the cultures that produce the textual objects being encoded, in that TEI projects are theoretical in their deliberate efforts to model document data (Ramsay and Rockwell 2012). We seek to encourage TEI-driven research on sex and gender by enhancing the *Guidelines*' expressiveness in these areas. Our revision of the *Guidelines* therefore provides examples but resists endorsing any single particular standard for specifying values for sex or gender. We recommend that projects encoding sex and/or gender explicitly state the theoretical groundwork for their ontological modeling, such that the encoding articulates a context-appropriate, informed, and thoughtful epistemology.

Gayle Rubin's influential theory of “sex/gender systems” informs some of our new language in the Guidelines “Names and Dates” chapter (Rubin 1975). While updating existing examples for encoding sex and introducing related examples for encoding gender, we mention the “sex/gender systems” concept to suggest that sex and gender may be *related*, such that a culture's perspective on biological sex gives rise to its notions of gender identity. Unexpectedly, we found ourselves confronting the *Guidelines*' prioritization of personhood in discussion of sex, likely stemming from the conflation of sex and gender in the current version of the *Guidelines*. In revising the technical specifications describing sex, we introduced the term “organism” to broaden the application of sex encoding. We leave it to our community to investigate the fluid concepts of gender and sex in their textual manifestations of personhood and biological life.

Encoding of cultural categories, when unquestioned, can entrench biases and do harm, a risk we must face in digital humanities generally. Yet we seek to make the TEI more expressive and adaptable for projects that complicate, question, and theorize sex and gender constructions. We look forward to working with the TEI community in hopes of continued revisions, examples, and theoretical document data modeling of sex and gender for future projects. In particular, we are eager to learn more from project customizations that “queer” the TEI and theorize about sexed and gendered cultural constructions, and we hope for a lively discussion at the TEI conference and beyond.

Works Cited

Ramsay, Stephen and Geoffrey Rockwell. 2012. “Developing Things: Notes toward an Epistemology of Building in the Digital Humanities.” *Debates in Digital Humanities*. University of Minnesota Press.
<https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfd1e/section/c733786e-5787-454e-8f12-e1b7a85cac72> .

Rubin, Gayle. 1975. “The Traffic in Women: Notes on the ‘Political Economy’ of Sex,” *Toward an Anthropology of Women*. Ed. Rayna R. Reiter. New York: Monthly Review Press, 157-210.

Sedgwick, Eve Kosovsky. 1990. "Introduction: Axiomatic." *Epistemology of the Closet*. University of California Press, 27-53.

Biographies

Elisa Beshero-Bondar is Professor of Digital Humanities and Program Chair of Digital Media, Arts, and Technology at Penn State Erie, The Behrend College. A scholar of gender and genre in 19th-century literature, she serves on the TEI Technical Council, leads the Digital Mitford project and works on machine-assisted collation and the XSLT production pipeline for the *Frankenstein Variorum*. Her adventures with markup technologies are documented on her development site at <https://newtfire.org>.

Helena Bermúdez Sabel is a postdoctoral researcher in the project *A world of possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language* (<https://woposs.unine.ch/>) hosted at the University of Neuchâtel (Switzerland). She has been a member of the TEI Technical Council since January 2021.

Raffaele (Raff) Viglianti is Senior Research Software Developer at the Maryland Institute for Technology in the Humanities, University of Maryland. His research is grounded in digital humanities and textual scholarship, where "text" includes musical notation. He researches innovative practices to model and publish textual sources as sustainable digital scholarly resources. He is a member of the Text Encoding Initiative Technical Council and the Technical Editor of the *Scholarly Editing* journal.

Janelle Jenstad is Professor of English at the University of Victoria. Her research interests are early modern drama, editorial praxis, and the geohumanities. She directs two TEI-powered and Endings-compliant projects: *The Map of Early Modern London* (MoEML) and *Linked Early Modern Drama Online* (LEMDO). She has been a member of the TEI Technical Council since January 2022. Her publications are listed at <https://janellejenstad.com/>.

Long Paper: Where is the Spanish in the TEI?: Insights on a Bilingual Community Survey

G. del Rio Riande¹, S. Allés-Torrent²

1: CONICET, Argentine Republic; 2: University of Miami, USA

Keywords: TEI, Spanish, Survey, Community, Geopolitics of Knowledge

Where is the Spanish in the TEI?: Insights on a Bilingual Community Survey

Susanna Allés-Torrent (UM)

Gimena del Rio Riande (CONICET)

Who can best define the interests and needs of a community? The members of the community itself.

"Communicating the Text Encoding Initiative to a Multilingual User Community" is a research project financed by the A. Mellon Foundation in which scholars from North and South America are generating linguistic, cultural and didactic situated educational materials to

improve XML-TEI encoding, editing and publication of Spanish texts. As it is well known, the TEI has as its primary goal the creation of a model that could semantically describe any text within any cultural heritage, any community, and any language. This goal is only feasible with a diverse and representative group of users that implement the guidelines and the encoding and publishing methodologies, and actively engage in discussions, events and activities. However, when we produce training materials, we often focus on the specific skills, capacities and tools we are trying to teach to individuals. And yet, identifying the community through shared narratives or culture is also crucial.

As part of the project activities, we prepared a bilingual survey (Spanish-English) aimed at inquiring t who uses or has used XML-TEI practices, and where and how they have been applied to Spanish humanistic texts. Bearing in mind that many digital scholarly edition projects of Spanish texts are carried out in Spanish-speaking and Anglophone institutions, we did not focus on a geographical survey, but on the use of XML at a global level. The survey ran between February and April 2022. It is an anonymous survey and consists of 22 questions. It received 104 responses, 77 in Spanish and 28 in English.

Some of the data that we will discuss in this short presentation aims at illustrating the significant differences regarding the organization of projects, collaboration, financing and use of TEI in master's and doctoral research. In broad terms, the survey allowed us to better understand not only the Spanish-speaking community that uses XML-TEI, but also to think of strategies that can contribute with more inclusive practices for scholars from less represented countries and in less favorable contexts inside the global TEI community. Last but not least, we believe the survey will be useful for designing actions that can support a wider range of modes of interaction and collaboration inside the global TEI community.

The survey is accessible at:

https://umiami.qualtrics.com/jfe/preview/SV_aWx84qH6cjh9Xf0?Q_CHL=preview&Q_SurveyVersionID=current

Session 2A – Long Papers – 11:30 - 13:00

Session 2A: Long Papers

Location: ARMB: 2.98

Chair: Elli Bleeker, Huygens Institute for the History of the Netherlands

Keywords: Herman Melville, genetic criticism, text analysis, R, XPath

Long Paper: Revision, Negation, and Incompleteness in Melville's *Billy Budd* Manuscript

C. Ohge

School of Advanced Study, University of London, United Kingdom

Revision, Negation, and Incompleteness in Melville's *Billy Budd* Manuscript Christopher Ohge (School of Advanced Study, University of London)

In 2019, John Bryant, Wyn Kelley, and I released a digital edition of Herman Melville's last work *Billy Budd, Sailor* (c. 1886–1891). This TEI-encoded edition required nearly 10 years to complete, mostly because this unfinished work survives in a complicated manuscript that demonstrates about 8 stages of composition. This born-digital 'fluid text' edition (Bryant 2002; Bryant et al 2019) consists of a diplomatic transcription of the manuscript, a 'base' (or clean) version of the manuscript, and a critical, annotated reading text generated from the base version. How could we effectively use all of the sophisticated descriptive markup of the manuscript transcription for critical purposes? What is missing is an analysis of this work's genesis that demonstrates the critical potentials of the encoding (Ohge 2021).

In this talk I will discuss my recent literary critical work based on text analyses of the TEI XML data of the *Billy Budd* manuscript. First I generated and visualised basic statistics of revision acts (counting additions, deletions, and substitutions) using XPath expressions and functions, and sorted the added and deleted words into separate lists. I then used the R programming language to perform analyses of the manuscript in comparison to Melville's oeuvre. With the XML2 library, I generated visualisations and statistical analyses of the XPath queries, including relative word frequencies of added and deleted words (Gries 2017; Jockers and Thalken 2020). With the TidyText library I turned all additions and deletions into a document term matrix and then into data frames in Hadley Wickham's 'tidy' format (Silge and Robinson 2022). In this format I can produce sentiment analyses: because of the TEI encoding, I can also perform these analyses on different versions of the text (i.e. before and after stages of revision).

Melville's deletions, negation words, negative sentiments, and incompleteness are conceived as a cluster of related textual phenomena that the TEI XML encoding helps to pinpoint. There are sheer frequencies relating to negation: firstly, that *Billy Budd* has more deletions in manuscript than additions, implying that Melville's tended to negate more than he added in composition, and secondly that frequencies of negation-words increased throughout his fictional work. (Negation-word results are generated through regular expression searches in the 'base' version of the text.) Although this trend drops off in the late poetry, *Billy Budd* has the highest number of negations in all of Melville's oeuvre.

That there are more deletions than additions in the *Billy Budd* manuscript reinforces its incompleteness. What the narrator called the 'ragged edges' of the story reflect not only

Melville's late tendency to rework words and ideas, but also to complicate the main characters of the novel (particularly Captain Vere) who represent justice in the story (Ohge 2021). I conclude by suggesting that this incompleteness is not only a metaphor in the text but a metaphor of the text of this tragic story.

References

- Bryant, John. 2002. *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. Ann Arbor: University of Michigan Press.
- Bryant, John, Wyn Kelley and Christopher Ohge. 2019. *Versions of Billy Budd, Sailor: A Fluid Text Edition*. Melville Electronic Library.
<https://melville.electroniclibrary.org/versions-of-billy-budd>.
- Gries, Stefan. 2017. *Quantitative Corpus Linguistics in R*. London: Routledge.
- Jockers, Matthew and Rosamond Thalken. 2020. *Text Analysis with R for Students of Literature*. 2nd edition. Cham, Switzerland: Springer.
- Ohge, Christopher. 2021. *Publishing Scholarly Editions: Archives, Computing, and Experience*. Cambridge: Cambridge University Press.
- Silge, Julia, and David Robinson. 2022. *Text Mining with R: A Tidy Approach*. O'Reilly.
<https://www.tidytextmining.com/index.html>.
-

Bio: Christopher Ohge is Senior Lecturer in Digital Approaches to Literature at the School of Advanced Study, University of London. His book *Publishing Scholarly Editions: Archives, Computing, and Experience* was published in 2021 by Cambridge University Press, and in 2022 he co-edited a special issue of *Textual Cultures* on creative-critical editing with Mathelinda Nabugodi. He also serves as the Associate Director of the Herman Melville Electronic Library.

Long Paper: “Un mar de sentimientos”. Sentiment analysis of TEI encoded Spanish periodicals using machine learning

L. Krusic¹, M. Scholger¹, E. Hobisch², Y. Völkl²

¹: Institute Centre of Information Modelling (Austrian Centre for Digital Humanities), University of Graz; ²: Technical University Graz

Keywords: digital editions, sentiment analysis, machine learning, literary analysis, corpus annotation

“Un mar de sentimientos”. Sentiment analysis of TEI encoded Spanish periodicals using machine learning

Sentiment analysis (SA), one of the most active research areas in NLP for over two decades, focuses on the automatic detection of sentiments, emotions and opinions in textual data (Liu, 2012). Recently, SA has also gained popularity in the field of Digital Humanities (Schmidt, Burghardt & Dennerlein, 2021; Rebora, Messerli & Herrmann 2022). This contribution presents the analysis of a TEI encoded digital scholarly edition (DSE) of Spanish periodicals using a machine learning approach for sentiment analysis as well as the re-implementation of the results into TEI for further retrieval and visualization.

This research is situated at the intersection of different projects. On the one hand, it builds on the project Distant Spectators (Scholger et al. 2019-2021), in which a sentiment analysis tool chain was developed (Koncar et al, 2022) for the investigation of 18th Century

Periodicals. These Spectators are a European journalistic phenomenon, propagating the social norms and values of Enlightenment by means of a characteristic narrative structure. On the other hand, this contribution constitutes a pre-study concerning the transition from the Spectator periodicals into epistolary novels in Spain – a topic which has remained without in depth analysis until now (Rueda 2001, 33, 181). The Spanish Spectator press consists of 690 issues from 23 periodicals and is available from the DSE *The Spectators in the international context* (Ertler et al. 2011-2021). The corpus is annotated following the standard of the TEI, considering both the logical text structure and narrative forms (such as reader's letters, self-portraits, or hetero-portraits) as well as subjects, places, persons, and intellectual works. Our approach to sentiment analysis was based on a manually created and computationally extended dictionary of words from the Spanish Spectator periodicals.

Currently, Spanish DH mostly relies on such dictionary-based tools (Moreno-Ortiz, 2017) and small corpora (Torres-Moreno & Moreno-Jiménez, 2020; Barbado et al 2021), while machine-learning approaches (García-Vega et al, 2019; Serrano et al, 2022) are the state-of-the-art in other domains. In this work, we compare our baseline, the dictionary-based SA of 23 Spanish periodicals (in total 690 issues), with a Python toolkit for Spanish SA, *pysentimiento* (Pérez et al, 2021). Moreover, we explore the variation of sentiments across selected narrative forms in the texts. To evaluate the performance of the current state-of-the-art transformer-based models provided by *pysentimiento* (Pérez et al, 2021) on our corpus, we conduct an annotation study to create a small evaluation corpus, such as LiSSS (Torres-Moreno & Moreno-Jiménez, 2020). Following the best practices for corpus annotation (Schmidt, Dangel, & Wolff, 2021), we include three expert annotators.

Usually, raw text is used for NLP tasks (such as SA). However, for a more detailed investigation, TEI encoded documents allow extraction (e.g. narrative forms) and exclusion (e.g. running heads) of certain text structures. When it comes to literary texts which are challenging for the task of SA due to their style, figures of speech and narrative forms, annotations can yield better classification results. Consequently, the results from the sentiment analysis can be re-implemented into the TEI encoding by using the @ana attribute on structural elements pointing to corresponding categories. This in turn allows exploration of the DSE through visualizations such as distribution of sentiments and sentiment development over time. The results can both facilitate a circular framework of the creation of a DSE as well as support the literary analysis and exploration of Spanish 18th Century Periodicals.

Bibliography

1. Barbado, A., Fresno, V., Riesco, Á. M., & Ros, S. (2022). DISCO PAL: Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels. *Language Resources and Evaluation*, 56(2), 501–542. <https://doi.org/10.1007/s10579-021-09557-1>
2. Ertler, K.-D., Fuchs A., Fischer-Pernkopf M., Hobisch E., Scholger M. & Völkl Y. (2011-2021). *The Spectators in the International Context*. <https://gams.uni-graz.at/spectators>.
3. García-Vega, M., Díaz-Galiano, M., García-Cumbreras, M., Plaza-Del-Arco, F.M., Montejo-Ráez, A., Zafra, S. M., Martínez-Cámara, E., Aguilar, C., Antonio, M., Cabezudo, S., Chiruzzo, L., Moctezuma, D. (2020). Overview of TASS 2020: Introducing Emotion Detection.

4. Koncar, P., Geiger, B. C.; Glatz, C.; Hobisch, E., Sarić, S., Scholger, M., Völkl, Y., Helic, D. (2022): A Sentiment Analysis Tool Chain for 18th Century Periodicals. In: Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, Ulrike Wuttke (Eds.): *Fabrikation von Erkenntnis. Experimente in den Digital Humanities*. Luxembourg. Zeitschrift für digitale Geisteswissenschaften und Melusina Press. 2022. DOI: <https://doi.org/10.26298/ezpg-wk34>.
5. Liu, B. (2012) *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers, Vermont, Australia.
6. Moreno-Ortiz, A. (2017). *Lingmotif: Sentiment Analysis for the Digital Humanities*. Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 73–76. <https://doi.org/10.18653/v1/E17-3019>.
7. Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and Social NLP tasks*. ArXiv:2106.09462 [Cs]. <http://arxiv.org/abs/2106.09462>.
8. Reborá, S., Messerli, T. C., & Herrmann, J. B. (2022): *Towards a Computational Study of German Book Reviews. A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis*. Conference poster. DHd2022. 7–11 March 2022.
9. Rueda, Ana (2001). *Cartas sin lacrar. La novela epistolar en la España Ilustrada 1789-1840*. Madrid: Iberoamericana.
10. Schmidt, T., Burghardt, M. & Dennerlein, K. (2021). *Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays*. In: Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, Ulrike Wuttke (Eds.): *Fabrikation von Erkenntnis. Experimente in den Digital Humanities*. Luxembourg. Melusina Press. DOI: <https://doi.org/10.26298/melusina.8f8w-y749-udlf>.
11. Schmidt, T., Dangel, J., & Wolff, C. (2021). *SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities*. ISI.
12. Scholger, M., Geiger, B., Hobisch, E., Koncar, P., Sarić, S., Völkl, Y.; Glatz, C. (2019-2021): *Distant Spectators (DiSpecs). Distant Reading for Periodicals of the Enlightenment*. <https://gams.uni-graz.at/dispecs>.
13. Serrano, A. V., Subies, G. G., Zamorano, H. M., Garcia, N. A., Samy, D., Sanchez, D. B., Sandoval, A. M., Nieto, M. G., & Jimenez, A. B. (2022). *RigoBERTa: A State-of-the-Art Language Model for Spanish*. ArXiv:2205.10233 [Cs]. <http://arxiv.org/abs/2205.10233>.
14. Torres-Moreno, J.-M., & Moreno-Jiménez, L.-G. (2020). *LiSSS: A toy corpus of Spanish Literary Sentences for Emotions detection*. ArXiv:2005.08223 [Cs]. <http://arxiv.org/abs/2005.08223>.

Session 2B – Long Papers – 11:30 - 13:00

Session 2B: Long Papers

Location: ARMB: 2.16

Chair: Hugh Cayless, Duke University

Long Paper: TEICollator: a semi-automatic TEI to TEI workflow

M. Gille Levenson

ENS Lyon, France

Keywords: collation, information transfer, ecdotics, materiality

TEICollator: a semi-automatic TEI to TEI workflow

Automated text comparison has been an area of interest for many years [Spadini 2016, Nury 2019]: tools such as CollateX allow automated text comparison, and even export to TEI. However, today there is no tool that allows, from transcripts encoded and structured in XML-TEI, to automate the collation of texts and to inject the produced apparatuses into the original files. Working in this way ensures that the contextual and structural information specific to each witness (structure, materiality, additions, deletions, line changes, etc) encoded in XML-TEI is not lost. In other words, there is a need of being able to work on textual differences without ignoring the individual, structural and material reality of each text or witness.

Furthermore, the increasing use of Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR) tools (see Kiessling 2019 for instance), which is interesting both in terms of speed of acquisition and of quality of the preserved information [Camps 2016], have consequences for the ecdotal methods: should we keep collating the text manually, when its acquisition has been done by the computer, knowing that XML is a common output of the latest HTR engines ?

My work focuses on a semi-automatic collation workflow. In this paper I will present a complete TEI to TEI processing chain that uses the Gothenburg model (tokenization, normalization, alignment, analysis/feedback, visualisation : see Spadini 2016) as a starting point, taking single TEI-encoded transcriptions to produce meaningful collated ones (see Camps *et al.* 2019), which keep the original structural information for each witness and identifies some of the most frequent textual features (omissions, homeotheleutons, simple transpositions) : I want to show how the TEI standard, the pivot format of this computational method, can be used to describe text as well as to process it.

I will outline the limitations and difficulties I face along the processing chain : in particular, the tool would be less efficient with important structural differences – a macro alignment (in paragraphs, for instance) is needed and, for the moment, has to be realized manually.

Finally, I will show how the transformation from TEI to LaTeX, maybe the most complex task, is fully part of the ecdotic chain, and contributes to produce meaning from the data: in this sense, my work is linked to the reflections carried out for several years on Digital Scholarly Editions [Pierazzo 2015; Pierazzo and Driscoll 2016] : I've decided to prefer the print/pdf format over a web interface, using the LaTeX Reledmac package [Rouquette 2022].

References

- Bleeker, Elli, Bram Buitendijk, R. Haentjens Dekker, et Astrid Kulsdom. « Including XML markup in the automated collation of literary text. » Prague, République Tchèque, 2018.
- Camps, Jean-Baptiste. “La Chanson d’Otinel : édition complète du corpus manuscrit et prolégomènes à l’édition critique,” 2016.
- Camps, Jean-Baptiste, Lucence Ing, and Elena Spadini. “Collating Medieval Vernacular Texts: Aligning Witnesses, Classifying Variants.” In DH2019 Digital Humanities Conference. Utrecht, 2019. Accessed November 7, 2020. <https://dh-abstracts.library.cmu.edu/works/10074>.
- Clérice, Thibault. “Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin.” Journal of Data Mining & Digital Humanities 2020 (April 7, 2020). Accessed September 3, 2020. <https://jdmhdh.episciences.org/6264/pdf>.
- CondorCompPhil/Falcon. For Alignment, Lemmatization and CollatioN. 2019. Accessed March 20, 2020. <https://github.com/CondorCompPhil/falcon>.
- Dekker, Ronald H., and Gregor Middell. “Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements.” Copenhagen, 2011.
- Jover, Francisco Gago, and Francisco Javier Pueyo Mena. “El Old Spanish Textual Archive, Diseño y Desarrollo de Un Corpus de Textos Medievales: El Corpus Textual.” Cuadernos del Instituto de Historia de la Lengua, no. 11 (2018): 165–209.
- Kiessling, Benjamin. “Kraken - an Universal Text Recognizer for the Humanities.” Utrecht, 2019. <https://dev.clariah.nl/files/dh2019/boa/0673.html>.
- Nury, Elisa. “Automated Collation and Digital Editions From Theory to Practice,” 2018. Accessed June 17, 2019. https://kclpure.kcl.ac.uk/portal/files/105283803/2018_Nury_Elisa_1337422_ethesis.pdf.
- Padró, Lluís, and Evgeny Stanilovsky. “FreeLing 3.0: Towards Wider Multilinguality.” In Proceedings of the Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey: ELRA, 2012.
- Pierazzo, Elena. Digital Scholarly Editing: Theories, Models and Methods. Ashgate Publishing, 2015.
- Pierazzo, Elena, and Matthew James Driscoll, eds. Digital Scholarly Editing: Theories and Practices. Open Book Publishers, 2016. Accessed June 15, 2021. <https://www.openbookpublishers.com/product/483>.
- Rouquette, Maïeul. Reledmac. Typeset Scholarly Editions with LATEX. TeX, 2021. Accessed January 16, 2022. <https://github.com/maieul/ledmac>.
- Spadini, Elena. « Studi sul Lancelot en prose ». PhD Thesis, Sapienza Università di Roma, 2016.

Long Paper: Back to analog: the added value of printing TEI editions

M. Kupreyev

Goethe Universität Frankfurt am Main, Germany

Keywords: digital edition, data quality assurance, XSL-FO, software test, PDF

Back to analog: the added value of printing TEI editions

Abstract: According to Sahle (2017) [1] digital editions are guided by a digital paradigm in their theory, method, and practice, and thus “cannot be given in print without significant loss of content and functionality”. In my talk I will touch upon the challenges of printing TEI XML datasets, but also highlight a useful diagnostic value of the PDF export for the data quality. PDF output, indeed, represents only a part of the encoded information, but it can play an essential role in data curation and quality assurance.

The “School of Salamanca” [2] project, jointly sponsored by the AdWL Mainz [3], MPI-LHLT [4] and Goethe-University Frankfurt [5], publishes the works of the jurists and theologians related to the University of Salamanca - the intellectual center of the Spanish monarchy during the 16th and 17th centuries. Based on a selected set of print editions we create a digital text corpus, which will include 116 works encoded in TEI XML. In addition, we also compose a historic dictionary of circa 300 essential terms, rendering the fundamental importance of the School of Salamanca for the early modern discourse about law, politics, religion, and ethics.

Our TEI XML data is controlled by the RNG schema and is exported to HTML and JSON IIF for web display [6]. Recently, a PDF printout option was added. Considering the complexity and the depth of annotation we decided to use the established XSL-FO technology, supported by a free Apache FOP processor integrated in the Oxygen Author workflow. Similar results might have been achieved with the CSS Paged Media Module or TEI Publisher. The PDF export highlighted issues which pertain to two ontologically different areas:

- Rendering XML elements in two-dimensional space of a PDF page.
- Semantic errors and inconsistencies in the XML encoding.

The issues of the first type refer, for example, to the representation of the marginal notes and their anchors, and to the correlation in pagination between XML, IIF and PDF. The problems of the second type include, for instance, different XML encoding of semantically identical chunks of information, which escaped the Schematron check-ups, but became visible with print layout.

PDF generation in the School of Salamanca was initially intended to be one of the export methods of the TEI data. It is now implemented early in the TEI production pipeline as a diagnostic tool, exposing the semantic and structural inconsistencies of the data, which can now be corrected before the final XML release. PDF production thus adheres to one of the principles of agile software testing, which states that capturing and eliminating defects in the early stages of research data life cycle is less time-consuming, less resource-intensive and less prone to collateral bugs (Crispin 2008) [7].

Biography:

Maxim N. Kupreyev is a software developer at the “School of Salamanca” project in Frankfurt am Main, where he is responsible for the production of XML-derived data formats such as IIF and PDF. Previously he was involved with the “Thesaurus Linguae Aegyptiae” project at BBAW [8] where he compiled and published the dataset of www.coptic-dictionary.org. He possesses 10 years of experience in software testing and 7 years in digital humanities, having obtained a

PhD degree in Egyptology and an ISTQB certificate [9]. His research interests include X-technologies, text processing and quality assurance of research data.

Bibliography:

- [1] Sahle, Patrick. 2017. "What is a Scholarly Digital Edition?" in *Digital Scholarly Editing*, edited by Matthew James Driscoll and Elena Pierazzo, 19-39. Cambridge: Open Book Publishers.
- [2] <https://www.salamanca.school/en/index.html> , accessed on 20.06.2022.
- [3] Akademie der Wissenschaften und der Literatur Mainz, <https://www.adwmainz.de/startseite.html>.
- [4] Max Planck Institute for Legal History and Legal Theory, <https://www.lhlt.mpg.de/en>.
- [5] Goethe Universität Frankfurt am Main, <https://www.goethe-university-frankfurt.de/en?locale=en>.
- [6] <https://blog.salamanca.school/de/2022/04/27/the-school-of-salamanca-text-workflow-from-the-early-modern-print-to-tei-all/>,
<https://blog.salamanca.school/de/2020/03/17/deutsch-entwicklung-der-webanwendung-v-2-0/> , accessed on 20.06.2022.
- [7] Crispin, Lisa. 2008. *Agile Testing: A Practical Guide for Testers and Agile Teams*. Addison-Wesley.
- [8] Berlin-Brandenburgische Akademie der Wissenschaften, <https://www.bbaw.de/en/>.
- [9] International Software Testing Qualification Board, <https://www.istqb.org/>

Long Paper: Encoding sonic devices: what is it good for?

M. Holmes

University of Victoria, Canada

Keywords: poetry, rhyme, sound

Encoding sonic devices: what is it good for?

The Digital Victorian Periodical Poetry project[1] has captured metadata and page-images for 15,548 poems from Victorian periodicals, and transcribed and encoded a representative sample of 2,150 poems. Our encoding captures rhyme and other sonic devices such as anaphora, epistrophe, and refrains. This presentation will describe our encoding practices and then discuss what useful outcomes can be gained from this undertaking. Although even TEI P1 specified both a rhyme attribute to capture rhyme-scheme and a rhyme element for "very detailed studies of rhyming" (TEI P1 P172)[2], and all significant TEI tutorials teach the encoding of rhyme (e.g. TEI by Example Module 4), it is difficult to find work which makes explicit use of TEI encoding of rhyme (let alone other sonic devices) in the analysis of English poetry.

Is manual encoding of rhyme still necessary? Chisholm & Robey noted back in 1995 that "much of the analysis which currently requires extensive manual markup will in due course be carried out by electronic means" (100), and much work has been devoted to the automated detection of rhyme (Kavanagh 2008; Kilner & Fitch 2017). However, these tools are not

completely successful, and in our own work, there is a consistent subset of cases which generate disagreement and discussion regarding type of rhyme, or even whether a rhyme is intended. We do make use of automated detection of anaphora and epistrophe, but only to generate suggestions for cases that might have been missed after the initial encoding has been done. We therefore believe that manually-curated encoding of sonic devices is a prerequisite for serious literary analysis which depends on that encoding.

Having invested in careful encoding of sonic devices, what are the potential uses for research? DVPP has begun by making rhyme-scheme discoverable and searchable in our search interface, and this is beginning to generate research questions. We can already test notions such as the claim that irregular rhyme-schemes were more frequently used as the century progressed; a table of the percentage of irregularly-rhymed poems in each decade in our collection (Appendix) shows only the weakest support for this claim.

In addition to tracing trends in poetic practice, and the construction of historical rhyme dictionaries, sonic device encoding might also be used for:

- Dialect detection. For example, our dataset includes a significant subset of poems written in Scots dialect, and others which may or may not be; for problem cases, where other factors such as poet and host publication suggest a dialect poem, but surface features are not persuasive, rhyme patterns may provide more evidence.
- Genre detection. Particular poetic genres, such as sonnets or ballads are characterized by formal structures which include rhyme-scheme.
- Bad poetry. We are particularly interested in the notion of what constitutes bad poetry, and our early work suggests that poetry which subjectively seems to be of poor quality also exhibits features such as monotonous rhyme-schemes and intrusive echoic devices.
- Authorship attribution.
- Diachronic sound-change.
- Historical rhyming dictionaries.

The presentation will present examples of cases where we have made use of our encoding to address questions such as these.

Notes:

[1] DVPP, <https://dvpp.uvic.ca/>

[2] See also Chisholm & Robey 1995.

Appendix:

TABLE: Prevalence of irregular rhyme in Victorian periodical poetry.

Decade	Percentage of irregularly-rhymed poems
1820s	4.1

1830s	3.4
1840s	3.4
1850s	1.1
1860s	8.6
1870s	3.6
1880s	3.7
1890s	5.4

References:

Chisholm, David, and David Robey. 1995. "Encoding Verse Texts." *Computers and the Humanities* Vol. 29, No. 2, *The Text Encoding Initiative: Background and Context* (1995), pp. 99-111. <https://www.jstor.org/stable/pdf/30200349.pdf>.

Kavanagh, F. 2008. "Analysis of a phonetic and rule based algorithm approach to determine rhyme categories and patterns in verse." MSc. Diss. *Computing Reports Technical Report No. 2007/23*. <http://computing-reports.open.ac.uk/2007/TR2007-23.pdf>.

Kilner, Kerry, and Kent Fitch. 2017. 'Searching for My Lady's Bonnet: discovering poetry in the National Library of Australia's newspapers database', *Digital Scholarship in the Humanities*, Volume 32, Issue suppl_1, April 2017, Pages i69–i83. <https://doi.org/10.1093/lc/fqw062>.

TEI P1 Guidelines for the Encoding and Interchange of Machine Readable Texts. 1990. Draft Version 1.1. Downloadable from <https://tei-c.org/Vault/Vault-GL.html>.

Terras, Melissa, Edward Vanhoutte, and Ron Van den Branden. *TEI by Example*. <https://teibyexample.org>. Last updated September 2020.

Session 3A – Long Papers – 14:30 - 16:00

Session 3A: Long Papers

Location: ARMB: 2.98

Chair: Gustavo Fernandez Riva, Universität Heidelberg

Long Paper: *Vocabularium Bruxellense*. Towards Quantitative Analysis of Medieval Lexicography

K. Nowak¹, I. Krawczyk¹, R. Alexandre²

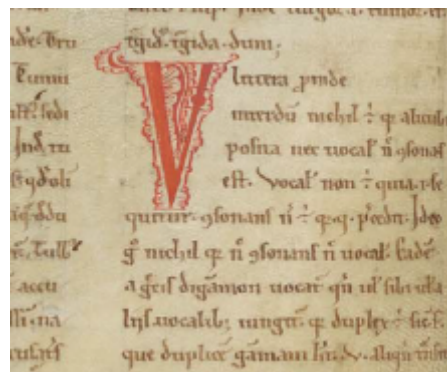
1: Institute of Polish Language (Polish Academy of Sciences), Poland; 2: Institut de recherche et d'histoire des textes, France

Keywords: Middle Ages, lexicography, glossary, quantitative analysis, Latin

Vocabularium Bruxellense. Towards Quantitative Analysis of Medieval Lexicography

The *Vocabularium Bruxellense* is a little-known example of medieval Latin lexicography (Weijers 1989). It has survived in a single manuscript dated to the 12th century and currently held at the Royal Library of Belgium in Brussels. In this paper we present the digital edition of the dictionary and the results of a quantitative study of its structure and content based on the TEI-conformant XML annotation.

First, we briefly discuss a number of annotation-related issues. For the most part, they result from the discrepancy between medieval and modern lexicographic practices which are accounted for in the 9th chapter of the TEI Guidelines (TEI Consortium) and the TEI Lex-0 recommendations (Tasovac et al. 2018). For example, a single paragraph of a manuscript may contain multiple dictionary entries which are etymologically or semantically related to the headword.



Ms. Bruxelles II 1049

Medieval glossaries are also less consistent in their use of descriptive devices. For instance, the dictionary definitions across the same work may greatly vary as to their form and content. As such, they require fine-grained annotation if the semantics of the TEI elements is not to be strained.

Second, we present the TEI Publisher-based digital edition of the *Vocabularium* (Reijnders et al. 2022). At the moment, it provides basic browsing and search functionalities, making the dictionary available to the general public for the first time since the Middle Ages.



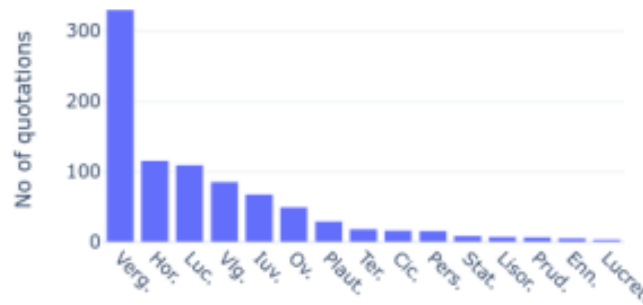
Digital edition of the Vocabularium Bruxellense

Thirdly, we demonstrate how the TEI-conformant annotation may enable a thorough quantitative analysis of the text which sheds a light on its place in a long tradition of medieval lexicography. We focus on two major aspects, namely the structure and the sources of the dictionary. As for the first, we present summary statistics of the almost 8,000 entries of the *Vocabularium*, expressed as a number of entries per letter and per physical page. We show that half of the entries are relatively short: a number among them contain only a one-word definition (usually, a synonym of the headword) and only 25% of entries contain 15 or more tokens.



Cumulative sum of the number of entries

Based on the TEI XML annotation of nearly 1200 quotes, we were able to make a number of points concerning the function of quotations in medieval lexicographic works which is hardly limited to attesting specific language use. We observe that quotations are not equally distributed across the dictionary, as they can be found in slightly more than 10% of the entries, whereas nearly 7,000 entries have no quotations at all. The quotes are usually relatively short with only 5% containing 10 or more words. Our analysis shows that the most quoted author is by a wide margin Virgil followed by Horace, Lucan, Juvenal, Ovid, Plautus, and Terence (19). Church Fathers and medieval authors are seldom quoted, we have also discovered only 86 explicit Bible quotations so far.



Most cited authors

In conclusion, we argue that systematic quantitative analyses of the existing editions of the medieval glossaries might provide useful insight into the development of this important part of the medieval written production.

References

- Reijnders, H. F., Krawczyk, Iwona, & Alexandre, Renaud. (2022). *Vocabularium Bruxellense. A Digital Edition (Version 1)*. Zenodo. <https://doi.org/10.5281/zenodo.6632046>.
- Tasovac, T., Romary, Laurent et al. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1*. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.
- TEI Consortium, eds. "9 Dictionaries." *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Version number]. [Last modified date]. TEI Consortium. <https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html> (10 June 2022).
- Olga Weijers. 1989. Lexicography in the Middle Ages. *Viator* 20: 139-53.

Long Paper: ISO MAF reloaded: new TEI serialization for an old ISO standard

P. Banski¹, L. Romary², A. Witt¹

1: IDS Mannheim, Germany; 2: INRIA, France

Keywords: standardization, morphology, morphosyntax, ISO, MAF, stand-off annotation

ISO MAF reloaded: new TEI serialization for an old ISO standard

Piotr Bański, IDS Mannheim

Laurent Romary, Inria

Andreas Witt, IDS Mannheim

{banski, witt} @ ids-mannheim.de, laurent.romary@inria.fr

Introduction

ISO Technical Committee TC 37, *Language and terminology*, Subcommittee SC 4, *Language resource management* (henceforth ISO TC 37/SC 4; see

<https://www.iso.org/committee/297592.html>) has been, for nearly 20 years now, the locus of

much work focusing on standardization of annotated language resources. Through the subcommittee's liaison with the TEI-C, many of the standards developed there use customizations of the TEI Guidelines for the purpose of serializing their data models. Such is the case of the feature structure standards (ISO 24610-1:2006, ISO 24610-2:2011), which together form chapter 18 (*Feature Structures*) of the Guidelines, as well as the standard on the transcription of the spoken language (ISO 24624:2016, reflected in ch. 8 - *Transcription of Speech*) or the Lexical Markup Framework (LMF) series, where ISO 24613-4:2021 maps the LMF model onto chapter 9 (*Dictionaries*) of the Guidelines.

The Morphosyntactic Annotation Framework (ISO 24611:2012; Clément and de la Clergerie, 2005) was initially published with its own serialization format, interwoven with suggestions on how its fragments can be rendered in the TEI. In a recent cyclic revision process, a decision was made to divide the standard in two parts, and to replace the legacy serialization format with a customization of the TEI that makes use of the recent developments in the Guidelines – crucially, the work on the standOff element and the work on the att.linguistic attribute class (cf. Bański and Romary, 2022). The proposed contribution reviews fragments of the revised standard and presents the TEI devices used to encode it. At the time of the conference, ISO/CD 24611-1 “Language resource management – Morphosyntactic annotation framework (MAF) – Part 1: Core model” will be under the Committee Draft ballot by the national committees mirroring ISO TC37 SC4.

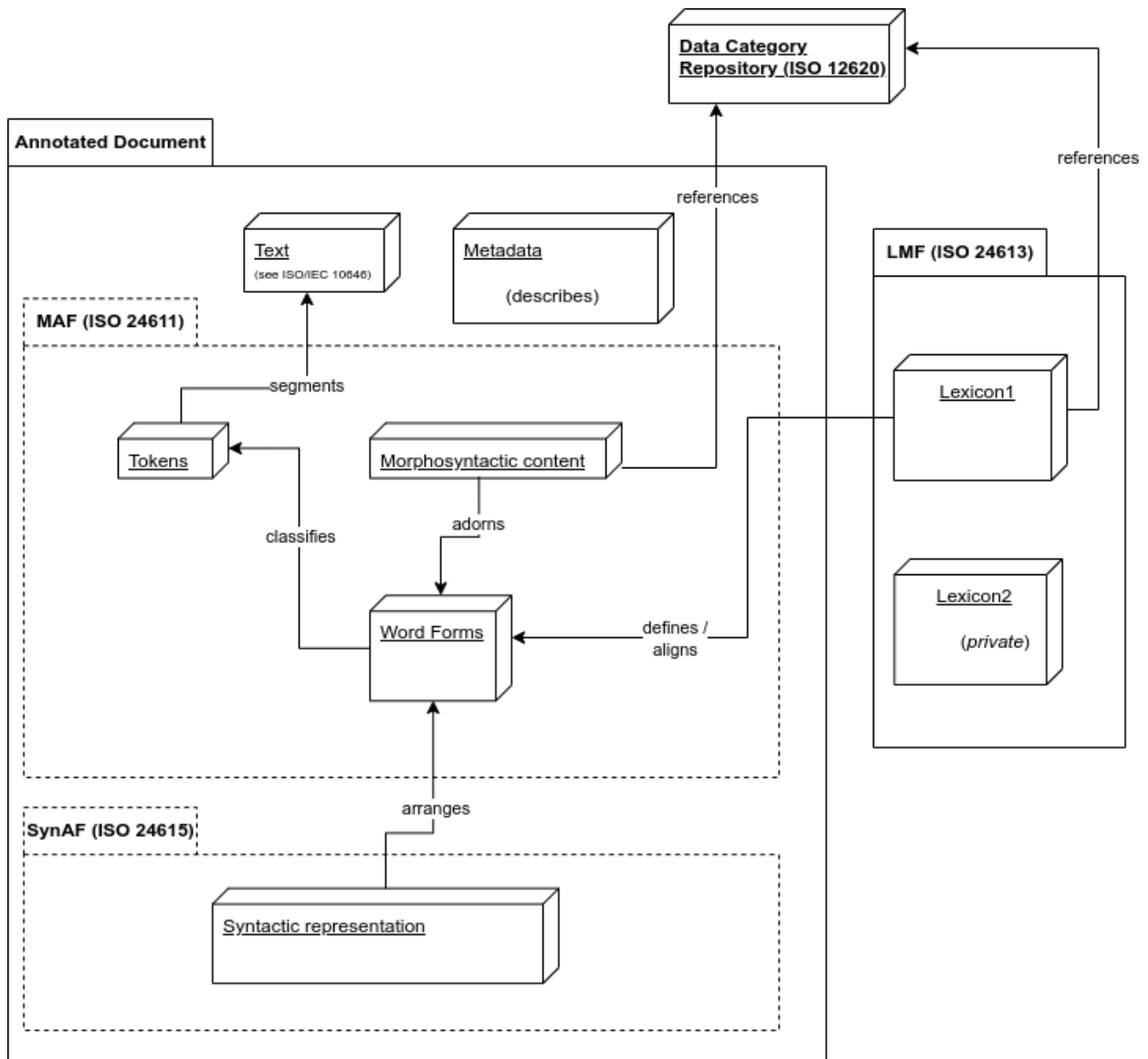
In what follows, we briefly outline the basic properties of the MAF data model and review selected examples of its serialization in the TEI.

ISO MAF data model: basic properties

The MAF data model assumes a crucial division in the levels of representation, known from classical approaches to linguistic morphology and from corpus linguistic studies, namely that between, on the one hand, the level of character sequences, identified by their unique position in the character stream forming the text under consideration and with string length as their further property and, on the other hand, the level of syntactic terminals, characterised by certain lexical and morphosyntactic properties and by non-unique relationships with objects on the former level. These two levels are referred to in the MAF standard as the level of tokens and the level of word-forms, respectively.

Apart from that distinction, ISO MAF strives to be as neutral with respect to linguistic formalisms as possible, attempting to provide a common ground for encoding nearly any approach, including overlapping or conflicting annotations.

Potential ambiguities in the tokenization and in the word-form composition, handled by a word-lattice mechanism, are the subject of the upcoming second part of the MAF standard, and we are therefore going to ignore them in the present submission.



The figure above presents the basic ecosystem of ISO MAF, in the context of an annotated document, which consists of (a) text proper, with minimal structural markup or without any markup at all, (b) formal metadata that describe the resource, and optionally (c) analytical metadata (annotations). Annotations that provide information on token-sized objects fall within the purview of ISO MAF, while annotations that describe how word-forms enter into syntactic combinations of various sorts are expected to conform to a family of standards known as ISO SynAF (Syntactic Annotation Framework). The figure omits the family of SemAF standards, which pertain to how syntactic structures and relations can be interpreted.

The document-based view is potentially complemented by the item-based lexica, described by ISO 24613 (Lexical Markup Framework, LMF). The morphosyntactic description of word forms, as well as the grammatical section of dictionary entries, can make reference to external taxonomies/ontologies, standardized by ISO 12620 (Data Category Repository, DCS).

ISO MAF serialization: selected examples

The TEI serialization of the ISO MAF attempts to reuse the existing pieces of the Guidelines, while adjusting them to the theoretical assumptions. Tokens are marked with <seg> elements as the most semantically neutral, while word-forms are encoded by means of

elements, which belong, among others, to the att.linguistic class (Bański, Haaf, Mueller, 2018) that enables the use of attributes defined for the purpose of encoding various kinds of word-level linguistic information (part of speech, morphosyntactic features, correspondence to dictionary entries, among others). The legacy approaches that, among others, use the elements <w> ('word') and <pc> ('punctuation character') for encoding tokens, thus mixing the levels of tokens and word forms as defined by ISO MAF, are not excluded by the standard, as long as they can be algorithmically converted into a representation using <seg> and .

Theory-neutrality is ensured by the optional use of the standOff element (see, among others, Bański et al., 2016), where annotations (also ones that conflict with one another) can be stored in separate layers of typed <annotationBlock> elements. This is illustrated by Example 1 in the Annex, where the text of the resource is located in the <text> element, and the standOff element consist of listAnnotation elements that correspond to a single unit of annotation (here: a sentence) and group separate layers of annotation, here: the tokenization layer with individual segments identified by character offsets (referencing inter-character points and counting from 0) and the word-form layer that references the tokens and identifies the part of speech of each syntactic terminal.

The paper will highlight some of the design choices and motivate the solutions adopted for them.

Annex: selected examples

```
<TEI>
(...)
<text>
<body>
<div>
<p xml:id="ex1.1">The victim's friends told the police that Krueger dove into the quarry and
never surfaced.</p>
</div>
</body>
</text>
<standOff>
<listAnnotation corresp="#ex1.1" xml:id="std.ex1.1">
<annotationBlock type="token" offsetBase="#ex1.1">
<seg xml:id="std.ex1.1.tok1" startPos="0" endPos="3"/>
<seg xml:id="std.ex1.1.tok2" startPos="4" endPos="10"/>
<seg xml:id="std.ex1.1.tok3" startPos="10" endPos="12"/>
<seg xml:id="std.ex1.1.tok4" startPos="13" endPos="20"/>
<seg xml:id="std.ex1.1.tok5" startPos="21" endPos="25"/>
<seg xml:id="std.ex1.1.tok6" startPos="26" endPos="29"/>
<seg xml:id="std.ex1.1.tok7" startPos="30" endPos="36"/>
<seg xml:id="std.ex1.1.tok8" startPos="37" endPos="41"/>
<seg xml:id="std.ex1.1.tok9" startPos="42" endPos="49"/>
<seg xml:id="std.ex1.1.tok10" startPos="50" endPos="54"/>
```

```

<seg xml:id="std.ex1.1.tok1" startPos="55" endPos="60" />
<seg xml:id="std.ex1.1.tok2" startPos="61" endPos="64" />
<seg xml:id="std.ex1.1.tok3" startPos="65" endPos="71" />
<seg xml:id="std.ex1.1.tok4" startPos="72" endPos="75" />
<seg xml:id="std.ex1.1.tok5" startPos="76" endPos="81" />
<seg xml:id="std.ex1.1.tok6" startPos="82" endPos="90" />
<seg xml:id="std.ex1.1.tok7" startPos="90" endPos="91" />
</annotationBlock>
<annotationBlock type="wordForm">
<span xml:id="std.ex1.1.wf1" target="#std.ex1.1tok1" pos="AT0" />
<span xml:id="std.ex1.1.wf2" target="#std.ex1.1tok2" pos="NN1" />
<span xml:id="std.ex1.1.wf3" target="#std.ex1.1tok3" pos="POS" />
<span xml:id="std.ex1.1.wf4" target="#std.ex1.1tok4" pos="NN2" />
<span xml:id="std.ex1.1.wf5" target="#std.ex1.1tok5" pos="VVD" />
<span xml:id="std.ex1.1.wf6" target="#std.ex1.1tok6" pos="AT0" />
<span xml:id="std.ex1.1.wf7" target="#std.ex1.1tok7" pos="NN2" />
<span xml:id="std.ex1.1.wf8" target="#std.ex1.1tok8" pos="CJT" />
<span xml:id="std.ex1.1.wf9" target="#std.ex1.1tok9" pos="VVB" />
<span xml:id="std.ex1.1.wf10" target="#std.ex1.1tok10" pos="NN1" />
<span xml:id="std.ex1.1.wf11" target="#std.ex1.1tok11" pos="NN1" />
<span xml:id="std.ex1.1.wf12" target="#std.ex1.1tok12" pos="AT0" />
<span xml:id="std.ex1.1.wf13" target="#std.ex1.1tok13" pos="NN1" />
<span xml:id="std.ex1.1.wf14" target="#std.ex1.1tok14" pos="CJC" />
<span xml:id="std.ex1.1.wf15" target="#std.ex1.1tok15" pos="AV0" />
<span xml:id="std.ex1.1.wf16" target="#std.ex1.1tok16" pos="SENT" />
<span xml:id="std.ex1.1.wf17" target="#std.ex1.1tok17" pos="PUN" />
</annotationBlock>
</listAnnotation>
</standOff>
</TEI>

```

Example 1. The use of the <standOff> element to ensure the ease of maintenance of layered annotations and to guarantee maximal theory-neutrality of the resulting description.

```

<p>
<s><seg xml:id="w1">I</seg>
<seg xml:id="w2">wanna</seg>
<seg xml:id="w3">put</seg>
<seg xml:id="w4">up</seg>
<seg xml:id="w5">new</seg>
<seg xml:id="w6">wallpaper</seg>
<seg>.</seg>
</s>
</p>

```



```
<spanGrp type="wordForm">
<span target="#w1" ana="#fs1" lemmaRef="#entry1" />
<span target="#w2" ana="#fs2" lemmaRef="#entry2" />
<span target="#w2" ana="#fs3" lemmaRef="#entry3" />
<span target="#w3 #w4" ana="#fs4" lemmaRef="#entry4" />
<span target="#w5" ana="#fs5" lemmaRef="#entry5" />
<span target="#w6" ana="#fs6" lemmaRef="#entry6" />
</spanGrp>
```

Example 2. Fragment of annotations added inside the <text> element, where <seg>ments are referenced by s encoding word-forms, which come with references to feature structures encoding comprehensive information on the given syntactic terminal, and with references to dictionary entries where more information can be sought.

References

ISO/WD 24611. Language resource management – Morphosyntactic annotation framework (MAF). Available from <https://www.iso.org/standard/79088.html>

Bański, Piotr; Gaiffe, Bertrand; Lopez, Patrice; Meoni, Simon; Romary, Laurent; Schmidt, Thomas; Stadler, Peter; Witt, Andreas. 2016. Wake up, standOff!. Presentation given at TEI Conference 2016, in Vienna, Austria. Available from <https://hal.inria.fr/hal-01374102v1>

Bański, Piotr; Haaf, Susanne; Mueller, Martin. 2018. Lightweight Grammatical Annotation in the TEI: New Perspectives. In: Nicoletta Calzolari (Conference chair) and Khalid Choukri and Christopher Cieri and Thierry Declerck and Sara Goggi and Koiti Hasida and Hitoshi Isahara and Bente Maegaard and Joseph Mariani and Hélène Mazo and Asuncion Moreno and Jan Odijk and Stelios Piperidis and Takenobu Tokunaga. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Available from <http://www.lrec-conf.org/proceedings/lrec2018/summaries/422.html>

Bański, Piotr and Laurent Romary. 2022. ISO/WD 24611 Morphosyntactic Annotation Framework (MAF) -- report on the revision process. Presentation given at the 11th meeting of ISO/TC 37/SC 4/WG 6 “Linguistic annotation” at DIN, Berlin. Available from https://www.researchgate.net/publication/362966073_ISO/WD_24611_Morphosyntactic_Annotation_Framework_MAF_report_on_the_revision_process

Clément, Lionel and Éric Villemonte de la Clergerie. 2005. MAF: a Morphosyntactic Annotation Framework. Available from https://www.researchgate.net/publication/228639144_MAF_a_Morphosyntactic_Annotation_Framework

Author Bios:

Piotr Bański works in the departments of Grammar and of Digital Linguistics at IDS Mannheim.

Among his current research interests are: standardization of language resources, corpus linguistics, electronic lexicography.

His projects and functions include:

- chairing the CLARIN Standards Committee,
- being co-convener of TEI LingSIG, former member of the TEI Council
- being expert of DIN, project leader of ISO CQLF, ISO MAF.

Laurent Romary works at INRIA, where he is the director for scientific information and culture.

Among his research interests is data modeling in the humanities.

His projects and functions include:

- chairing ISO committee TC37 (language and terminology)
- being a convener of ISO/TC 37/SC 4/WG 4 (lexical resources)
- being co-chair of the DARIAH lexical resources working group, which has edited the TEI Lex0 specification
- being a former member of the TEI Council and the TEI Board.

Andreas Witt works at IDS Mannheim, where he heads the Digital Linguistics department.

Among his current research interests is digital research infrastructure for the humanities.

His projects and functions include:

- chairing the Consortia Assembly of the National Research Data Infrastructure Germany in Germany (since April 2022)
- being a member of the CLARIN Board of Directors (2018-2021)
- being co-convener of TEI LingSIG
- convening ISO TC37 SC4 WG6.

Long Paper: TEI Modelling of the Lexicographic Data in the DARIAH-PL Project

K. Nowak, D. Mika, W. Łukasik

Institute of Polish Language (Polish Academy of Sciences), Poland

Keywords: lexicography, dictionaries, semantic web

Krzysztof Nowak, Dorota Mika, Wojciech Łukasik

TEI Modelling of Lexicographic Data in the DARIAH-PL Project

The main goal of the “DARIAH-PL Digital Research Infrastructure for the Arts and Humanities” project is building the Dariah.lab infrastructure, which would allow for sharing and integrated access to digital resources and data from various fields of the humanities and arts. Among numerous tasks that the Institute of Polish Language, Polish Academy of Sciences coordinates, we are working towards the integration of our lexicographic data with the LLOD resources (Chiarcos et al. 2012). The essential step of this task is to convert the raw text of a dozen of paper-born dictionaries into TEI-compliant XML format (TEI Consortium).

In this paper we would like to outline the main issues involved in TEI XML modelling of these heterogeneous lexicographic data.

In the first part, we give a brief overview of the formal and content features of the dictionaries. For the most part, they are multi-volume works developed between the 1950s and 2010s with the research community in mind, and as such they are rich in information and structurally complex. They cover diachronic development (from medieval Polish and Latin to present-day Polish) and the functional variation of Polish (general language vs. dialects, proper names).

Vispron	3	Volmerus
Vispron* : Obiit) ... Vispron ca 1265 LMP s. 704.		
Vitalis cf. Witalis(r) .		
Vitelus lat.: Per manus magnifici Erasmi Vitellii, secretarii (Alexandri ... ducis Lithuaniae) 1499 UPL 71 s. 120; Per manus magistri Erasmi Vitellii, praepositi et canonici Vilmensis, secretarii nostri (Alexandri, ducis Lithuaniae) (1499) WIn 464 s. 545 (cf. Erasmus Stanislai Ciolek de Cracovia 1485 AS 1 s. 270). — Cf. I. Cholek.		
Vithmarus cf. Witmar .		
Vithwald* : Nobili Alberto (dot.) Vithwald 1486 Paw s. 111.		
1. Vitoldus cf. 1. Witolt A.		
2. Vitoldus cf. 2. Witolt.		
Vitrarius lat.: Ad molendinum Vitrarri XV (1330) KW 1114.		
Vitus cf. Wit .		
Viwuch* : Ioh. Viwuch 1396 RmK II s. 146.		
Vladimirus cf. Wiedźmir 2.		
Vladislaides : Mescio Vladislaides ... Vladislaidem Boleslaum XIV (ca 1194—1205) Koff s. 397, 405; Casimirus Vladislaides lagilo 1463/1464 (1450) MPH III s. 120 (cf. Serenissimi principis, domini Vladislai ... regis Poloniae ... et ... filiorum suorum ... Vladislai et Kasimiri 1432 UPL 57 s. 83); Duces Plozenses Vladislaides, scilicet Semovitus et Vladislau 1463/1464 (1462) MPH III s. 121 (cf. Vladislau, Masoviae et Plocensis princeps ... filius duobus Semovito et Vladislao (1455) DOp XIV s. 217); Vladislaides (I pro Vladislaides) Meskonem et Boleslaum XIV—XV (XIII) Bog s. 351 (cf. Boleslaus, filius Vladislai cum fratre suo Mescone XV (1153) MPH II s. 873); Anna, filia Gedymini ... consors Kasimir, regis Poloniae Vladislaidis dicti Lechicensis XV (XIII—XV; sub a. 1335) MPH III s. 199 (cf. Kasimirus, filius regis Poloniae Wladyslay dicti Loctek duxit uxorem, Annam nomine ... filiam Gedymini, ducis Lituanorum ca 1341 (sub a. 1325) MPH II s. 854).		
Vladislau cf. Władysław A. 2.		
Vladislavita : Magistro Iohanne de Wladyslaw, decretorum doctore ... doctor Vladislavita 1402 ARex 1517; In causa doctoris Vladislavita (I) 1493 ARex 1254; Doctorem Vladislaviam ... prepositus Collegij Canoniarum, doctor Vladislavita 1455 ARex 1740.		
Vlantsczak* : Iacobus Vlantsczak (I) dedit pro ½ orlo 1378 Cdm s. 94.		
Vlascides cf. Włoszczowie B. 2.		
Vleck cf. Fick.		
Vlodislau cf. Władysław A. 1.		
Vlodes cf. Włoszczowie B. 1.		
Vlodesis cf. Włoszczowie B. 1.		
Vlodonisa cf. Włostewa B.		
Vnarco* : Vnarco subcamerario 1242 FG 954 w. 9.		
Vncusteterissa* fem.: Dedit Vncusteterissa quartale 1½ grossum 1369 Cdm s. 15.		
Vobis lat.: Domina Katherina de Brzunico, uxor Laurentii dicti Vobis 1413 Cae 235; Laurentius dictus Vobis de Brzunico nomine procuratorio Katherine, uxoris sue 1415 Cae 309; Super Laurentium dictum Vobis 1415 Cae 330.		
Vobistigwaszd* : Magister Iacobus de Casimira, medicinae doctor insignis, cognominatus Vobistigwaszd, loquutionibus enim suis frequenter verbum istud "vobistigwaszd" apponebat 1470—1480 DAB III s. 120.		
Vocabula lat.: Mathias Vocabula, procurator generis Stephani Pogorsky 1490 Hcl II 4367; Honorabilis dominus Mathias Vocabula ... dominus Mathias Vocabula 1497 Hcl II 4538.		
Vogelingsang cf. Fagellingsang 1.		
Vogll cf. Fag(Oct) 2.		
Vogllingsang cf. Fagllingsang.		
Vogllingsang cf. Fagllingsang 2.		
Vogulo* : Sophia, Vogulo, Iohannes ca 1266 LMP s. 705.		
Vohclur* : Vohclur, Ida et Craynu sorores (I) nortre ca 1265 LMP s. 683.		
Volkone* : A Volkone, scriptore regis 1345—1363 MVat II 277 s. 452.		
Voiniciensis cf. Wojciez(yski) B.		
Volkane cf. Folkane 1.		
Volkmarus cf. Folkmar 1.		
Volmarus cf. Folkmar 2.		
Volmerus cf. Folkmar 3.		

Dictionary of Old Polish Personal Names

BAGROWNIK 'robotnik obsługujący pogłębiarkę': Bagrownik robę na bagnach *Kasz S VII 5*. TG

BAGRÓWKI 'ryby — węprze płotki': Bagrówki Łąk kosiń. TG

Bagteryjka zob. **BATERYJKA**

I. BAGUN 'żołędź u zacierzą': Bagun Orasna w-tar. TG

II. BAGUN *Forma*: Bayun *Gietrzwałd* olsz [N Troki, Wilno ZSER] *MAGP XI s 63*; bahun *Rogów* [Poniewież ZSER] ju.

Znaczenie: 'roślina — bagno (Ledum palustre)': *Gietrzwałd* olsz [N Troki, Wilno ZSER] ju; *Rogów* [Poniewież ZSER] ju. TG

BAHAJMA 'o szlowskiu: niezłara': Taką bahajma — cywóek niedorożony *Cergowa kros*. TG

BAHAJMO 'o czymś bezwartościowym': To takie bahajmo *Przędzel wś*. TG

Bahastwo zob. **BAGASTWO**

Bahaża zob. **BAGAŻA**

Baher zob. **BAGIER**

Bahla zob. **BAGLA**

BAHNIATKA 'agrest': Bahniatka *Sromowce Wi w-tar MPEJ I 67*; ~ 'dziśki agrest': Bahniatki, to just dziśki agrest i różne inne jagody przynosi *Szczawiewa w-tar KŁap 15*. TG

Bahniak zob. **BAGNIUK**

Bahniukowy zob. **BAGNIUKOWY**

Bahno zob. **BAGNO**

BAHIRA 'obwód koła w wodzie nie kuty': Bahra *Gładkie-Zakopane MAGP I s 73*. TG

BAIRO *Forma*: Bagro *Falejówka szosce AJPP 139*.

Znaczenie: 'całość lub część drewnianego obrotu koła umocowanego na sprzączkach; szosow': Nla. sprzyrag [kółu] bayra okute razem *Głodówka* [Orasna Czel] *ZNUJ OLI 23*; Przednie kolo ma pyńg bayra, a zadnie ma soś. Bayra [luśeniowa, a sprzy] tyz. [laśniowa *Sucha Góra* [Orasna Czel] ju 114; F k'wiośo je soś bayrał, dwanás sprzy *Nowoś* [Orasna Czel] ju 37; Sprzy [...] jaśniowa, a bayra sám blukowe. Na bayrak sám [obrynosie] sie'ozne *Pogóra* [Orasna Czel]

ju 58; *Jablonków* [Ciechyn Czel] *PIJP I ss 43 s 58*; *Jaworzynka ciecz*; *Oznaczénica* [Ciechyn Czel] *AJS II s 6*; *Kasz* [Spicz Czel] *AJPP 139*; *Frydman w-tar ZNUJ CCLXXVIII 49*; *N-tar RWP XVII 25*; [st, nie uż] *Brzema-Litacz w-ąd PPodogr 164*; *Falejówka szosce AJPP 139*; *Lubomierz doś PE II 357*. TG

Bahrować zob. **BAGROWAĆ**

BAHRY bhp 'chwast o niebieskich kwiatach rosnący w czołu': Bahry *Dubica wślad*. TG

Bahun zob. **II. BAGUN**

BAI I. w funkcji ekspresywnej I. *szamaśia* całą *szponicóś*: *Smoga* — to je taki to lepkaru, to uż tjeśnia puśo, ni jyny smrek baji *Lipółka Komeralna* [Ciechyn Czel] *AJS II s 59*; a. w potępieniu 'oczyszczenie': Bai tak! Bai za tak! *Sucha* [Ciechyn Czel]; *Mim pjiśóć* baji pfié *Kozakowice ciecz*; *G Sobieszowice* [Ciechyn Czel] *Kel II 128*; b. w pytaniu 'czy': *Rozmo'łalyśóć* Baji nie! *Zyczya gar*.

2. podkreślenie *jedyn* z *szlonów* *szponicóś* a. *szmarocząc* *uwagę* na *coś* 'na przykład, oto': *Niekiedy* baby na *krew* nie *chciały* [miski] *dzierżoś*, *nimiały* ku *tymu* *notury*. *Baj* moja *nieboga* *matka*, *ta* *nigdy* na *krew* nie *dzierżała* *Sucha Średnia* [Ciechyn Czel] *Zwrot 83 s 11*; *Szli* *strasznie* *długo*, *jako* *baj* z *Jablonkowa* *do* *Pragi*, *a* *można* *jeszcze* *dali* *Kozaczyszka* [Ciechyn Czel] *PiegKoz 21*; *Strasnie* *winter* *tam* *fućo* *y* *śóegrym* *plu* *jako* *baj* *śiso* *Puśóć* *ciecz*; b. *ucydnisuje* a. *śóóś* 'aś': *Wypily* *gorzoly* *had* *cztery* *garney* [pén] *Strzelecki* *prud* *Piećniśi III 54*; *ś. intensywność* *działania*, *niezwykłość* *czepoś* 'nawet': *Za* *łocy* *jakim* *darymniie* *pop'żuzyvou*, *baj* *wykudlip* [wytarguś za *włosy*] *Puśóć* *ciecz* *NY I 5*; *Płynico*, *to* *uż* *baj* *małny* *rypśne* *niefi* [do *młóćcia*], *nó* *Ścozyciel* [Ciechyn Czel] *BMJP XIV 77*; c. *ogranicza* *tylko*, *jedynie*: *To* *uż* *nie* *nie* *wort* — *tu* *baj* *śfiniim* *a* *krowóm* *dajóm* *Sucha* [Ciechyn Czel]; d. *przy* *dopowiedzeniu*, *zdanis* *używanym* 'zreśta': *Dostoloch* *za* *kułly*, *nawoś* *był* *baj* *przy-*

Dictionary of Polish Dialects

- 35 | **Zły** *formy*: n. sg. m. zły *Gn* 14b. 173a. 181b, etc. etc.; f. zła *Gn* 12a, 1449 *R* XXV 165, ca 1450 *PF* IV 568, etc.; neutr. zle *Gn* 183b, *Fl* i *Put* 77, 10. 90, 10, etc. etc.; ~ g. sg. m. złego *Fl* 70, 5, *Fl* i *Put* 42, 1. 100, 5, etc.; f. zle *BZ* Lev 21, 7, *Skarga* *Wroc* w. 58; złej *Fl* i *Put* 118, 101, 1449 *R* XXV 164, XV *med.* *R*

Dictionary of Old Polish

On a practical level, this means that, first, substantial effort had to be put into optimizing the quality of the OCR output. Since, except for grobid-dictionaries (Khemakhem et al. 2018), there are no tools at the moment that would enable easy conversion of lexicographic data, the subsequent phase of structuring of dictionary text had to be applied on a *per resource* basis.

TEI XML annotation has three main goals. First, it is a means of preserving the textuality of dictionaries which make heavy use of formatting conventions to convey information and employ a complex system of text-based internal cross-references. Second, TEI modelling aims at a better understanding of each resource and its explicit description. The analysis is performed by lexicographers who may, however, come from a lexicographic tradition different from the one embodied in a particular dictionary, and thus need to make their interpretation of the dictionary text explicit. Regardless, this way we may also detect and correct editorial inconsistencies, which are natural for collective works developed over many years. Third, the annotated text is meant to be the input used in the alignment and linking tasks, it is therefore crucial that functionally equivalent structures are annotated in a systematic and coherent way. As we plan to provide an integrated access to the dictionaries, the TEI XML representation is also where the first phase of data reconciliation takes place. It does not only concern the structural units of a typical dictionary entry, such as <sense/> or <form/>, but also mapping between units of analytical language the dictionaries employ, such as labels, bibliographic reference system etc.

References

- Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer, Heidelberg, p. 201-216.
- Mohamed Khemakhem, Axel Herold, Laurent Romary. 2018. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In: GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan. 2018.
- TEI Consortium, eds. "9 Dictionaries." *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Version number]. [Last modified date]. TEI Consortium.
<https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html> (10 June 2022).

Session 3B – Panel: Notes from the DEPCHA Field and Beyond: TEI/XML/RDF for Accounting Records – 14:30 - 16:00

Session 3B: Notes from the DEPCHA Field and Beyond: TEI/XML/RDF for Accounting Records

Location: ARMB: 2.16

Chair: Syd Bauman, Northeastern University

Panel: Notes from the DEPCHA Field and Beyond: TEI/XML/RDF for Accounting Records

K. Tomasek¹, O. Bullock¹, L. Hermsen², R. Walker², N. Kokaze³

1: Wheaton College Massachusetts, United States of America; 2: Rochester Institute of Technology, United States of America; 3: Chiba University, Japan

Keywords: accounts, accounting, DEPCHA, bookkeeping ontology

Notes from the DEPCHA Field and Beyond: TEI/XML/RDF for Accounting Records Session Proposal—Panel of Short Papers

Abstract:

The short papers in this session focus on questions that arise in the process of editing manuscript account books. Some of these questions result from the “messiness” of accounting practices in contrast to the “rationality” of accounting principles; others arise from efforts to reflect in the markup social and economic relationships beyond those imagined in Chapter 14 of the P5 TEI Guidelines, “Tables, Formulae, Graphics, and Notated Music.” The Bookkeeping Ontology developed by Christopher Pollin for the Digital Edition Publishing Cooperative for Historical Accounts (DEPCHA) in the Graz Asset Management System (GAMS) extends the potential of TEI/XML using RDF.

In “Wheaton Accounts and the DEPCHA Model,” Tomasek introduces the GAMS environment, Pollin’s ontology, and the relationship of Bullock’s summer 2022 work on a Wheaton account book to previous Wheaton accounts that students transcribed and marked up in summer 2016. If time allows, Tomasek will offer examples of a draft taxonomy based on *Historical Statistics of the United States*, a resource for economic history originally published by the U.S. Bureau of the Census. The goal of the taxonomy is to develop additional semantic markup to supplement Pollin’s Bookkeeping Ontology.

In “Operating Centre Mills,” Bullock focuses on what she learned transcribing and marking up information about the people, materials, and machines used to produce cotton batting at Centre Mills, a textile mill in Norton, Massachusetts, in 1847-48. The general ledger for this enterprise includes store accounts, production records, and tracking of materials used to run the mill. Entries that reflect the costs of mill operation show sources of raw cotton, daily use of materials, and payments for wages and board for a small labor force. Examples in the paper demonstrate flexible use of the <measure> element.

“Wages and Hours,” Hermsen and Walker’s paper, emerges from their work on a digital scholarly edition of account books of William Townsend & Sons, Printers, Stationers, and Account Book Manufacturers, Sheffield UK (1830-1910). Volume 3, “Business Guide and Works Manual,” speaks both to book history and to cultural observations about unionization, gender roles, and credit/debit accounting.

The financial accounts in this collection record the cost of doing business: expense for insurance and utilities, price for bookbinding materials, and rates for labor by position of foreman, women sewers, and apprentices. Townsend records numerous figures for labor expense throughout the manuscript. These figures provide information regarding the total hours worked per week and at what rate before and after unionization, according to union law, and with overtime. The ledger then offers additional insight by listing wage expenses by department and total wages for individual managers (including back wages disbursed to widows) by year. The financial records in this volume are managed by the author’s adaptation rather than bookkeeping systematization. Accounting for goods and services is logged in ambiguous tabular form, often in the margins of pages with unrelated jottings, and frequent in-text page references to nearly indecipherable price keys, or cross-references to the firm’s journals and ledgers.

Naoki’s paper, “Stakeholders in the British Ship-Breaking Industry,” develops a set of methods to analyse structured data of historical financial records, taking a disbursement ledger of Thomas W. Ward, the largest British shipbreaker in the twentieth century, as an example. That ledger is held by the Marine Technology Special Collection at Newcastle University, UK. The academic contribution of this research is to critically examine the possibilities and limitations of DEPCCHA, the ongoing digital humanities approach for semantic datafication of historical financial records with the TEI and RDF, mainly developed by scholars in the United States and Austria, and to present an original argument in British maritime history, which is to visualise a part of the overall structure of the British shipbreaking industry.

Development of DEPCCHA was supported by a joint initiative of the National Historic Publications and Records Commission at the National Archives and Records Administration in the United States and the Andrew W. Mellon Foundation.

Bios:

Kathryn Tomasek is Professor of History at Wheaton College. She has been working on TEI for account books since 2009, and she was PI for the DEPCCHA planning award in 2018. She chaired the TEI Board between 2018 and 2021.

Olivia Bullock is a senior Creative Writing major at Wheaton College who studies intersectional identities in literature and history.

Lisa Hermsen is Professor and Caroline Werner Gannett Endowed Chair in the College of Liberal Arts at Rochester Institute of Technology.

Rebecca Walker, Digital Humanities Librarian, coordinates large-scale DH projects and supports classroom digital initiatives in the College of Liberal Arts at Rochester Institute of Technology.

Naoki Kokaze is an Assistant Professor at Chiba University, where he leads the design and implementation of DH-related lectures in the government-funded humanities' graduate education program conducted in collaboration with several Japanese universities. He is a PhD candidate in History at the University of Tokyo, writing his doctoral dissertation focusing on the social, economic, and diplomatic aspects of the disposal of obsolete British Royal Navy's warships from the mid-nineteenth century through the 1920s.

Posters Slam and Session – 16:30 - 18:00

Poster Slam, Session, and Reception

Location: **ARMB: King's Hall**

Chair: Syd Bauman, Northeastern University

The Poster Slam and Session will start with a 1 minute - 1 slide presentation by all poster presenters summarising their poster and why you should come see it.

There will be an informal drinks and nibbles reception during the poster session.

Where a poster image was submitted for the Virtual Poster Session which followed on 22 September 2022 at 1pm BST, then this image has been included following the poster abstract.

Poster: The QhoD project: A resource on Habsburg-Ottoman diplomatic exchange

S. Kurz, M. Mayer, Y. Yilmaz

Austrian Academy of Sciences, Austria

Keywords: Early modern history, Ottoman, Edition

Having started as a cross-disciplinary source-editing project (Early modern history, Ottoman studies) in 2020, the *Digitale Edition von Quellen zur habsburgisch-osmanischen Diplomatie 1500–1918* (QhoD) project has recently become accessible to public with a rich collection of TEI-based source editions concerning diplomatic exchanges between the Ottoman and Habsburg empires. The project is editing sources related to tangible diplomatic missions, focussing on the grand embassies. The physical documents are predominantly housed in the Austrian and Turkish State Archives, but also in libraries, museums and private archives. Provided that the archives agree, QhoD presents digital facsimile data to the sources. To date, documents pertaining to four missions have been edited in sub-projects, with regular additions of newly edited documents scheduled for the next months and years.

Unique features of the project:

- QhoD is editing sources from both sides (Habsburg and Ottoman archives), giving complimentary views; Ottoman sources are translated into English (German language sources will shortly include abstracts in English)
- diversity of source genres (e.g. letters, contracts, travelogues, descriptions and depictions of cultural artifacts in LIDO; protocol register entries, *seyahatnâme* and *sefâretnâme* (traditions of Ottoman itineraries, the latter specific to embassies), newspapers, etc.)
- openness to outside collaboration (bring your TEI data!)

For Ottoman sources, QhoD is adhering to the *İslam Ansiklopedisi Transkripsiyon Alfabesi* transcription rules (Arabopersian to Latin transliteration). Transcriptions are aided by using Transkribus HTR mainly for German language sources, with ventures into Ottoman HTR together with other projects. Named entity data is curated in a shared instance of the Austrian Prosopographical Information System (APIS), aligned to GND identifiers and serialized as `tei:standOff`.

By the time of writing, <https://qhod.net> features

- by language: 60 German, 42 Ottoman language documents
- by genre: 60 letters, 20 protocol-register entries, 16 official records, 5 artifacts, 4 travelogues, 4 reports, 3 instructions
- by embassy/timeframe: 16 sources related to correspondence between Maximilian II and Selim II (1566–1574); 31 sources on Rudolf Schmid zu Schwarzenhorn's internuntiate (1649); 61 sources on the mutual grand embassies of Damian Hugo von Virmont and Ibrahim Pasha (1719–1720)

The poster will describe those sources and the TEI-infused reasoning behind their edition, as well as the technical implementation, which uses the GAMS repository software to archive and disseminate data.

QhoD uses state-of-the-art TEI/XML technology to improve availability of archival material essential for understanding centuries of mutual relations between two large imperial entities.

Bibliography

- Gürkan, Emrah Safa. "Mediating Boundaries: Mediterranean Go-Betweens and Cross-Confessional Diplomacy in Constantinople, 1560-1600." *Journal of Early Modern History* 19, no. 2-3 (2015): 107-28.
- Işıksel, Güneş. "Ottoman Diplomacy." In *The Encyclopedia of Diplomacy*, edited by Gordon Martel, ? New York: John Wiley & Sons Inc, 2018.
- Pešalj, Jovan. "Monitoring Migrations: The Habsburg-Ottoman Border in the Eighteenth Century." Unpublished Ph.D. dissertation, Leiden University, 2019.
- Radway, Robyn Dora. "Vernacular Diplomacy in Central Europe: Statesmen and Soldiers between the Habsburg and Ottoman Empires, 1543-1593." Princeton University, 2017.
- Sowerby, Tracey A. "Early Modern Diplomatic History." *History Compass* 14, no. 9 (September 2016): 441-56.
- Sowerby, Tracey A., and Christopher Markiewicz, eds. *Diplomatic Cultures at the Ottoman Court, c.1500-1630*. New York: Routledge, 2021.
- Strohmeyer, Arno. 2013a. "Die Theatralität interkulturellen Friedens: Damian Hugo von Virmont als Kaiserlicher Großbotschafter an der Hohen Pforte (1719/20)." In *Frieden und Friedenssicherung in der Frühen Neuzeit. Das Heilige Römische Reich und Europa. Festschrift für Maximilian Lanzinner zum 65. Geburtstag*, 413-438.
- Strohmeyer, Arno. 2013b. "Kategorisierungleistungen und Denkschemata in diplomatischer Kommunikation: Johann Rudolf Schmid zum Schwarzenhorn als kaiserlicher Resident an der Hohen Pforte (1629-1643)." In *Politische Kommunikation zwischen Imperien. Der diplomatische Aktionsraum Südost- und Osteuropa*, 21-29.
- Strohmeyer, Arno. 2014. "Krieg und Frieden in den habsburgisch-osmanischen Beziehungen in der Frühen Neuzeit." In *Die Türkei, der deutsche Sprachraum und Europa. Multidisziplinäre Annäherungen und Zugänge*, hg. von Reiner Arntz, Michael Gehler, 31-50.
- Yılmaz, Yasir. 2017. "Nebulous Ottomans vs. Good Old Habsburgs: a historiographical comparison." *Austrian History Yearbook* 48: 173-190.

- Yılmaz, Yasir. 2021. "From Nemçe to Avusturya: Ottoman Appellations for Austria." In Was heißt Österreich?: Überlegungen zum Feld der Austrian Studies im 21. Jahrhundert, hg. von Sieglinde Klettenhammer und Kurt Scharr, 80-97.



QhoD

Digital Scholarly Edition of Habsburg-Ottoman Diplomatic Sources

1500–1918

About the QhoD project

QhoD is a digital scholarly edition of Ottoman and Habsburg diplomatic sources from the 16th to the 19th century. The project is a collaboration between the University of Vienna and the University of Salzburg. It aims to provide a comprehensive digital edition of these sources, including transcriptions, translations, and original images. The project is funded by the Austrian Science Fund FWF and the Austrian Ministry of Education, Science and Research.

Sources are edited in sub-projects






Unique features of QhoD

- Open access to the original sources and their transcriptions.
- Open access to the original sources and their translations.
- Open access to the original sources and their images.
- Open access to the original sources and their metadata.

Editorial principles of the project

All texts of diplomatic correspondence are presented in their original language. The original language is Latin, German, or Ottoman Turkish. The original language is indicated by the language code in the XML. The original language is indicated by the language code in the XML. The original language is indicated by the language code in the XML.

Example: The Grand Embassy of Damjan Hugo von

Example: The Grand Embassy of Damjan Hugo von
Vittoriano Bevilacqua Zonta (1792-1802)

For the first time, throughout the history of the Republic of Venice, a Venetian ambassador was appointed to the Ottoman Empire. The ambassador was Damjan Hugo von Vittoriano Bevilacqua Zonta. He was appointed in 1792 and served until 1802. His mission was to establish diplomatic relations between the Republic of Venice and the Ottoman Empire.



Digital Workflow

The digital workflow of the project is as follows: 1. Identification of sources. 2. Acquisition of sources. 3. Transcription of sources. 4. Translation of sources. 5. Metadata creation. 6. Publication of sources.

Participating

The project is a collaboration between the University of Vienna and the University of Salzburg. It is supported by the Austrian Science Fund FWF and the Austrian Ministry of Education, Science and Research.

Team and Contacts

Project Director: Sieglinde Klettenhammer
Project Manager: Kurt Scharr
Project Assistant: Yasir Yılmaz

Bibliography

Yılmaz, Yasir. 2021. "From Nemçe to Avusturya: Ottoman Appellations for Austria." In Was heißt Österreich?: Überlegungen zum Feld der Austrian Studies im 21. Jahrhundert, hg. von Sieglinde Klettenhammer und Kurt Scharr, 80-97.














Bibliography

Yılmaz, Yasir. 2021. "From Nemçe to Avusturya: Ottoman Appellations for Austria." In Was heißt Österreich?: Überlegungen zum Feld der Austrian Studies im 21. Jahrhundert, hg. von Sieglinde Klettenhammer und Kurt Scharr, 80-97.

Poster: Building a digital infrastructure for the edition and analysis of historical travelogues

S. Balck

IOS Regensburg, Germany

Keywords: Semantic Web, Mobility Studies, Travelogues

Sandra Balck* (balck@ios-regensburg.de), Anna Ananieva (ananieva@ios-regensburg.de), Hermann Beyer-Thoma (hermann@beyer-thoma.de), Ingo Frank (frank@ios-regensburg.de), Jacob Möhrke (moehrke@ios-regensburg.de), Corwin Schnell (schnell@ios-regensburg.de), Leibniz Institute for East and Southeast European Studies (IOS) Regensburg, Germany

Abstract

Editions of historical travelogues are often based on document modelling with TEI. These editions lack the expressiveness and flexibility to answer complex research questions regarding travel (sub-)events (departure, arrival, etc.) and travel observations (visited public places, habits of people, etc.). To enable the analysis and visualisation of the material, our project is working on a digital edition combining TEI with semantic web and linked data technologies.

We will use an annotated critical text edition of the unpublished records of Franz Xaver Bronner's (1758 - 1850) journey from Aarau, via St. Petersburg, to the university in Kazan (1810) and his way back (1817) via Moscow, Lviv and Vienna as a case study. The aim will be to develop a modularly expandable digital research infrastructure, which will support digital transcription, annotation and visualisation of travelogues.

In the preliminary stages of the project, the first part (the outward journey) of Franz Xaver Bronner's travelogue manuscript has been transcribed with Transkribus. Training modules were developed based on manually transcribed texts, these are to be used for the semi-automatic transcription of other related texts. People, places, travel and other events were annotated with XML markup elements using TEI. In the next step, ontology design patterns for travelogues and itineraries will be developed. This includes a new annotation scheme for linking the TEI annotated text passages to associated database entries. The edition will enable the visualisations of textual information and contextual data.

This new effort goes beyond existing projects, such as the Hellespont project (Mambrini 2016) or the Semantic Blumenbach edition (Wettlaufer 2015) due to the development and application of ontology design patterns. Our approach makes more explicit data modelling possible and enables for further analysis and visualisation.

Works Cited

Mambrini, Francesco (2016). "Treebanking in the world of Thucydides. Linguistic annotation for the Hellespont Project". In: Digital Humanities Quarterly 10 (2).
<http://www.digitalhumanities.org/dhq/vol/10/2/000251/000251.html>

Wettlaufer, Jörg et al. (2015). "Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science". In: Digital Scholarship in the Humanities 30 (1), pp. i187-i198. <https://doi.org/10.1093/llc/fqv047>

Building a Digital Infrastructure for the Edition and Analysis of Historical Travelogues

Sandra Balck* (balck@ios-regensburg.de), Anna Ananieva, Hermann Beyer-Thoma, Ingo Frank, Jacob Mohrke, Corwin Schnell*(schnell@ios-regensburg.de)

Editions of historical travelogues are often based on document modelling with TEI. These editions lack the expressiveness and flexibility to answer complex research questions regarding travel (sub-)events (departure, arrival, etc.) and travel observations (visited public places, habits of people, etc.). To enable the analysis and visualisation of the material, our project is working on a digital edition combining TEI with semantic web and linked data technologies.



Figure 1: Example for a transcription via Transkribus lite



Figure 2: Experimental initial data model of the database



Figure 3: Screenshot of the first text annotations via TEI

Use Case: F. X. Bronner

We will use an annotated critical text edition of the unpublished records of Franz Xaver Bronner's (1758 - 1850) journey from Aarau, via St. Petersburg, to the university in Kazan (1810) and his way back (1817) via Moscow, Lviv and Vienna as a case study. The aim will be to develop a modularly expandable digital research infrastructure, which will support digital transcription, annotation, and visualisation of travelogues.

Preliminary work

In the preliminary stages of the project, the first part (the outward journey) of Franz Xaver Bronner's travelogue manuscript has been transcribed with Transkribus. Training modules were developed based on manually transcribed texts; these are to be used for the semi-automatic transcription of other related texts. People, places, travel, and other events were annotated with TEI according to our annotation guidelines. In the next step, ontology design patterns for travelogues and itineraries will be developed.

Preview: Ontology Design Patterns

We use the TEI subset DTABf (Haaf et al. 2015) on the text side and will develop an ontology-based text enrichment and editing workflow on the data side. Ontology design patterns will be iteratively constructed and used for classification of travel (sub-)events (departure, arrival, etc.) and travel observations (e.g. visited public places, habits of people, etc.). Where TEI will be used for text markup, CRM will be used for enrichment of the text with knowledge made explicit and stored in a database.

This new effort goes beyond existing projects, such as the Hesperion project (Mambrini 2016) or the Semantic Blumenbach edition (Wettlaufer 2015) due to the development and application of ontology design patterns. Our approach makes more explicit data modelling possible and enables for further analysis and visualisation.

Work cited: Mambrini, Francesco (2016). "Freebanking in the world of Thucydides. Linguistic annotation for the Hesperion Project". In: Digital Humanities Quarterly 10 (2). // Wettlaufer, Jörg et al. (2015). "Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science". In: Digital Scholarship in the Humanities 30 (1), pp. i187-i198.



Figure 4: Reconstruction of F. X. Bronner's travel route created by Hermann Beyer-Thoma (2022)



DEHisRe - Digitale Edition Historischer Reiseberichte
dehisre.ios-regensburg.de



Poster: TEI and Scholarly Digital Editions: how to make philological data easier to retrieve and elaborate

C. Martignano

University of Florence, Italy

Keywords: scholarly digital editions, conceptual model, digital philology, textual criticism, text modeling

Abstract

In the past few decades the number of TEI-encoded scholarly digital editions (SDEs) has risen significantly, which means that a big amount of philologically edited data is now available in a machine-readable form. One could try to apply computational approaches, in order to further study the linguistic data, the information about the textual transmission, etc. contained in multiple TEI-encoded digital editions. The problem is that retrieving philological data through different TEI-encoded SDEs is not that simple.

Every TEI-encoded edition has its own markup model, designed to respond to the philological requirements of that particular edition. The TEI guidelines, for example, show how the @type attribute can be used with the <rdg> element to distinguish between different types of variants. However, every edition may have its own set of possible values, beyond “orthographic” and “substantive”, to markup a wider range of phenomena of the textual transmission. For this reason, it is difficult to identify the same types of data through different digital editions unambiguously.

A possible way to simplify the retrieval of philological information from multiple digital editions is to link them to a same model that is able to represent SDEs on a more abstract level. This abstract model could be formalised as an ontology. Then inside different TEI-encoded editions it would be possible to add a further markup layer that binds each abstract component to the corresponding class of the ontology.

There are already some ontologies that were created for representing scholarly edited texts [14] and, more specifically, the critical apparatus [11]. My goal is to use these existing models, as well as the TEI guidelines, to analyse different scholarly editions (both digital and printed) and identify all the abstract components of a scholarly edition. The final goal of this preliminary research is to lay a theoretical foundation for an abstract model that can help make philological data more visible and easier to retrieve.

Bibliography

1. ‘12 Critical Apparatus - The TEI Guidelines’. n.d. Accessed 18 August 2022.
<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html>.
2. Barabucci, Gioele, Elena Spadini, and Magdalena Turska. 2017. ‘Data vs. Presentation. What Is the Core of a Scholarly Digital Edition?’ In *Advances in Digital Scholarly Editing*, 37–46. Sidestone Press.
https://serval.unil.ch/notice/serval:BIB_09C6C598108A.
3. Ciotti, Fabio, and Francesca Tomasi. 2016. ‘Formal Ontologies, Linked Data, and TEI Semantics’. *Journal of the Text Encoding Initiative*, no. Issue 9 (September).
<https://doi.org/10.4000/jtei.1480>.

4. Doerr, Martin. 2003. 'The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata'. *AI Magazine* 24 (3): 75–75. <https://doi.org/10.1609/aimag.v24i3.1720>.
5. Eide, Øyvind. 2014. 'Ontologies, Data Modeling, and TEI'. *Journal of the Text Encoding Initiative*, no. Issue 8 (December). <https://doi.org/10.4000/jtei.1191>.
6. Mancinelli, Tiziana, and Elena Pierazzo. 2020. *Che cos'è un'edizione scientifica digitale*. Carocci.
7. Pierazzo, Elena. 2014. *Digital Scholarly Editing: Theories, Models and Methods*. <http://hal.univ-grenoble-alpes.fr/hal-01182162>.
8. ———. 2019. 'What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter'. *International Journal of Digital Humanities*, May. <https://doi.org/10.1007/s42803-019-00019-3>.
9. Roelli, Philipp, ed. 2020. *Handbook of Stemmataology: History, Methodology, Digital Approaches*. De Gruyter. <https://doi.org/10.1515/9783110684384>.
10. Sahle, Patrick. 2016. 'What Is a Scholarly Digital Edition?' In *Digital Scholarly Editing*, edited by Matthew James Driscoll and Elena Pierazzo, 1st ed., 4:19–40. Theories and Practices. Open Book Publishers. <https://www.jstor.org/stable/j.ctt1fzhh6v.6>.
11. 'The Critical Apparatus Ontology (CAO)'. n.d. Accessed 18 August 2022. <https://fgiovannetti.github.io/cao/>.
12. Van Zundert, Joris, and Peter Boot. 2011. 'The Digital Edition 2.0 and the Digital Library: Services, Not Resources'. *Digitale Edition Und Forschungsbibliothek (Bibliothek Und Wissenschaft)* 44: 141–52.
13. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.
14. N.d. Accessed 19 October 2020. <http://e-editiones.ch/ontology/scholarly-editing>.

Biography:

Chiara Martignano is a research fellow at the Department of Literature and Philosophy of the University of Florence and a PhD candidate in Philology and Textual Criticism at the University of Siena. Her research focuses on scholarly digital editing and the development of web apps for digital editions. She is currently working as the digital humanist of the ERC-funded European Ars Nova project. She has collaborated with the Edition Visualization Technology project as a web developer.

TEI and Scholarly Digital Editions

How to make philological data easier to retrieve and visualise

Chiara Martignano
chiara.martignano@unifi.it
University of Florence, University of Siena

THE IDEA

1. In the past few decades the number of TEI-encoded scholarly digital editions (SDEs) has risen significantly, as a consequence a big amount of philologically edited data is now available in a machine-readable form. One could try to apply computational approaches, in order to further study the linguistic data, the information about textual transmission, etc. contained in multiple TEI-encoded digital editions. The problem is that **retrieving philological data through different TEI-encoded SDEs is not that simple.**

2. Every TEI-encoded edition has its **own markup model**, designed to respond to the philological requirements of that particular edition. For this reason, it is difficult to **identify the same types of data** through different digital editions **unambiguously.**

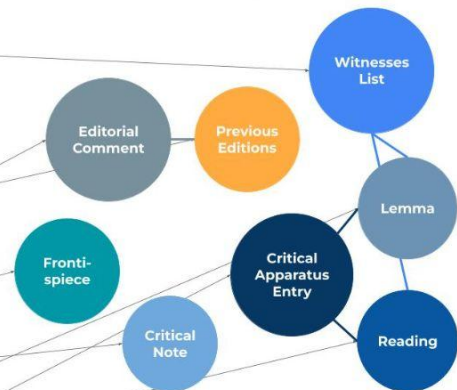
A MARKUP LAYER

```
<sourceDesc>
  <listWitnes>
    <witness xml:id="B">
      <desc>
        ...
      </desc>
    </witness>
  </listWitnes>
</sourceDesc>

<text xml:id="rebus_siculis_carmen" subtype="editorial_text" type="verse" xml:id="text_carmen">
  <front>
    <titlePage>
      <titlePart>
        <h1 rend="italic">De rebus Siculis carmen</h1> di Pietro da Eboli
      </titlePart>
    </titlePage>
    <p>Questa edizione critica è stata curata da Fulvio Delle Donne: poiché il ms. è quasi certamente idiografo, si è deciso di rispettarne la grafia. ...</p>
    <fw place="top-left" type="edizioni">
      <h1 rend="bold">Principali precedenti edizioni</h1>
      <listBibl xml:id="Edizioni" next="#studi">
        <biblstruct xml:id="Egell1746">
          <monogr>...</monogr>
        </biblstruct>
      </listBibl>
    </front>
    <div type="frontespizio" n="6" xml:id="frontesiz">
      <fw place="top-middle" type="frontespizio" rend="bold">PETRUS DE EBULO</fw>
      </div>
    <div type="Liber" n="1" xml:id="pde.lib.1">
      <div type="Particula" n="1" xml:id="ing.99v">
        <note n="1">Qui il verso, come anche in altre occasioni, è lasciato incompiuto dall'autore: forse volontariamente, per istituzione di <opertione ref="Virg">Virgilio</opertione>
      </note>
      <div type="facsimile" n="1" xml:id="facsimile">
        <img alt="Facsimile of a manuscript line" data-bbox="312 480 559 534"/>
      </div>
    </div>
  </text>
```

Example taken from the digital critical edition of *Petri de Ebuli De rebus siculis carmen*, edited by Fulvio Delle Donne.

AN ABSTRACT MODEL FORMALISED AS AN ONTOLOGY



3. A possible way to simplify the retrieval of philological information from multiple digital editions is to **link them to a same model** that is able to **represent SDEs on a more abstract level.** This abstract model could be formalised as an **ontology.** Then inside different TEI-encoded editions it would be possible to add a further **markup layer** that binds each abstract component to the corresponding class of the ontology.

POSSIBLE APPLICATIONS



Testing the **completeness** and **expressiveness** of existing ontologies for the description of scholarly digital editions. See [2].



Queries across different TEI-encoded editions, with RDF and SPARQL. See [1], [4].



Easier **development of general purpose tools** for visualising SDEs thanks to a common abstract model. See [5].

CURRENT STATE OF THE RESEARCH

4. There are already some ontologies that were created for representing **scholarly edited texts** and, more specifically, **the critical apparatus.** My goal is to **use these existing models**, as well as the **TEI guidelines**, to **analyse different scholarly editions** (both digital and printed) and **identify all the abstract components** of a scholarly edition. The final goal of this preliminary research is to lay a **theoretical foundation** for an abstract model that can help make philological data more visible and easier to retrieve.

THEORETICAL FOUNDATION FOR THE ABSTRACT MODEL

STUDY OF EXISTING MODELS



ANALYSIS OF THE COMPONENTS OF SCHOLARLY EDITIONS



REFERENCES

1. Ciotti, Fabio, and Francesca Tomasi. 2016. 'Formal Ontologies, Linked Data, and TEI Semantics'. *Journal of the Text Encoding Initiative*, no. Issue 9 (September). <https://doi.org/10.4000/tei.1480>.
2. Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi. 2019. 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini'. *Umanistica Digitale*, no. 7 (December). <https://doi.org/10.6092/issn.2532-8816/909>.
3. Doerr, Martin. 2003. 'The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata'. *AI Magazine* 24 (3): 75-75. <https://doi.org/10.1609/aimag.v24i3.1720>.
4. Eide, Øyvind. 2014. 'Ontologies, Data Modeling, and TEI'. *Journal of the Text Encoding Initiative*, no. Issue 8 (December). <https://doi.org/10.4000/tei.1191>.
5. Martignano, Chiara. 2021. 'Un modello concettuale per favorire lo sviluppo e il riutilizzo di app per edizioni digitali'. *Umanistica Digitale*, no. 10 (September): 71-88. <https://doi.org/10.6092/issn.2532-8816/2620>.

Poster: Between Data and Interface, Building a Digital Library for Spanish Chapbooks with TEI-Publisher

E. Leblanc, P. Jacsont

Keywords: Spanish literature, Digital library, Services, TEI-Publisher

Between Data and Interface, Building a Digital Library for Spanish Chapbooks with TEI-Publisher

Elina Leblanc, Pauline Jacsont

This poster will present the project Untangling the cordel (2020-2023) and its experimentations with TEI-Publisher to develop a digital library (DL) that aims at studying and promoting the Geneva collection of Spanish chapbooks (Leblanc and Carta 2021). Intended for a wide audience and sold in the streets, chapbooks recount fictitious or real events as well as songs, dramas, or religious writings. Although their contents are varied, they are characterised by their editorial form, i.e. small texts (4 to 8 pages), in in-quarto, arranged in columns and decorated with woodcuts. The interest in chapbooks ranges from literature to art and book history, sociology, linguistics, or musicology. This diversity reflects the hybridity of chapbooks, at the frontier between document, text, image, and orality (Botrel 2001; Gomis and Botrel 2019, 127–30).



Figure 1: Examples of Spanish chapbooks (From left to right: D. Juan de Serrallonga, *El Abanico*, Barcelona, [s.d.]; *El cantor de las hermosas*, T. Gaspar, Barcelona, [s.d.]; Antonio Narvaez y Rosaura, José María Moreno, Carmona, 1858; *Despertador espiritual*, José María Moreno, Carmona, 1858)

An editorial workflow based on XML-TEI to display our corpus online was devised. After transcribing texts with HTR tools, they were 1) converted the transcriptions in XML-TEI via XSLT, 2) stored them in eXist-DB, and 3) published them with TEI-Publisher. Images of the documents are displayed with IIIF. Through this workflow, the DL can offer services that stress different aspects of chapbooks: parallel consultation of transcription and facsimiles, comparison of documents with Mirador, full-text and thematic searches, woodcut catalogue, etc.

Working with TEI-Publisher has influenced the way we think about our XML-TEI model. If the choices we have made are mainly driven by data, it appears that part of them

have been influenced by the functionalities we wanted to implement, such as the addition of image links or keywords. Thus, our ODD reflects not only the nature of our documents but also the DL services. In this context, the use of TEI-Publisher invites us to reconsider a strict distinction between “data over interface” and “interface over data” (Dillen 2018), as data and interface are here mutually influenced.

Bibliography:

- Botrel, Jean-François. 2001. ‘El Género de Cordel’. In *Palabras Para El Pueblo. I. Aproximación General a La Literatura de Cordel*, edited by Luis Díaz G. Viana, 41–69. Madrid: CSIC.
http://www.cervantesvirtual.com/obra-visor/el-gnero-de-cordel-0/html/0133d94a-82b2-11df-acc7-002185ce6064_7.html#I_0_.
- Dillen, Wout. 2018. ‘The Editor in the Interface: Guiding the User through Texts and Images’. In *Digital Scholarly Editions as Interfaces*, by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider, 35–59. Norderstedt: BoD - Book on Demand.
- Gomis, Juan, and Jean-François Botrel. 2019. “‘Literatura De Cordel’ From A Transnational Perspective. New Horizons For An Old Field Of Study’. In *Crossing Borders, Crossing Cultures*, edited by Massimo Rospoche, Jeroen Salman, and Hannu Salmi, 127–42. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110643541-008>.
- Leblanc, Elina, and Constance Carta. 2021. ‘Le projet « Démêler le cordel » : une bibliothèque numérique pour l’étude de la littérature éphémère espagnole du XIXe siècle’. In *Humanistica 2021*, Rennes, 10–12 mai 2021, 100–101.
<https://hal.archives-ouvertes.fr/hal-03526522>.

Biographies:

Elina Leblanc is a postdoctoral researcher at the Spanish Unit of the University of Geneva for the project *Untangling the cordel* (2020-2023). She oversees the editorial workflow and the development of the digital library. In 2019, she obtained her PhD in Digital Humanities from the University of Grenoble (France): “Enriched Digital Libraries: Users, Services, Interfaces”, supervised by the professors Elena Pierazzo and Hervé Blanchon. In her thesis, she explored the notion of participative services, users, and interfaces in the context of the *Fonte Gaia* project.

After studying modern and classical literatures at the University of Lyon 3 and Paris Sorbonne, Pauline Jacsont defended her master’s thesis at the University of Neuchâtel in 2021, which focused on an XML-TEI digital edition of Latin tragedies with a textometric exploitation. She is pursuing her training in Digital Humanities at Geneva University. In parallel, she collaborates on different research projects such as *Woposs*, *FoNDUE*, and *Untangling the cordel* where she offers her expertise regarding text encoding. niversity of Geneva, France

BETWEEN DATA AND INTERFACE

Building a Digital Library for Spanish Chapbooks with TEI-Publisher

The corpus

Spanish chapbooks are characterized by:

- Their **editorial format**: small texts (4 to 8 pages), in-4°, arranged in columns and decorated with woodcuts;
- The **variety of their contents**: real events, fictitious stories, songs, poems, plays, religious writings, etc.

The Geneva's collection contains **915** Spanish chapbooks:

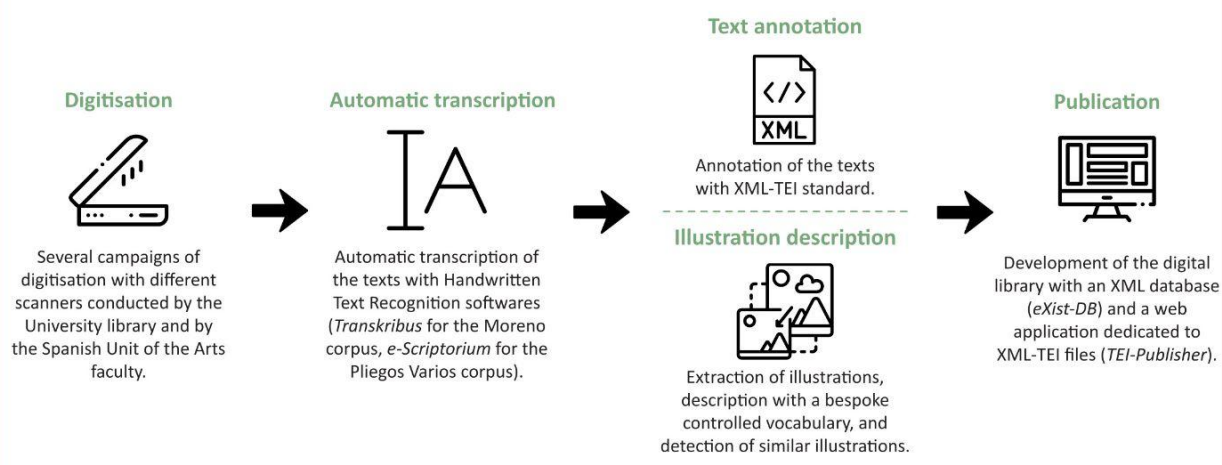
- 539 documents originated from the **major printing centers** in Spain (Madrid, Barcelona, Valencia, Palma de Mallorca);
- 376 documents printed by the **same printer**, in Carmona.

The output



The method

Development of a digital library with **digital scholarly editions** and a **catalog of illustrations**.



Data and interface

An ODD that reflects the **nature of the documents** and the **digital library's services**:

- A **data-driven approach**, in most cases;
- Some elements **influenced by the services** we want to implement and by the way *TEI-Publisher* works (Ex.: facsimiles, keywords, links, etc.).

→ **Mutual influence** of data and interface, that invite us to reconsider the strict distinction traditionally made between them.

An example

Displaying the facsimile alongside the transcription

What we originally plan

Using `<facsimile>` to list the images and to link them to the text with an attribute `@facs` at a page level.

What we currently do

Using the attribute `@facs` alone, with for value the URI of the image on the IIF server of the Geneva University.

Credits

Project manager: Constance Carta (Assistant professor).

Research associates: Luana Bermúdez and Belinda Palacios (Senior lecturers).

Digital humanities: Elina Leblanc (Postdoctoral researcher) and Pauline Jacsont (Research associate).



Scan me to see the website!



Poster: oXbytei and oXbytao. A Stack of Configurable oXygen Frameworks

C. Lück

Universität Münster, Germany

Keywords: software, editors, oxygen, frameworks, annotations

oXbytei and oXbytao

A Stack of Configurable oXygen Frameworks

Christian Lück

Until recently, adapting author mode frameworks for the oXygen XML editor was rather limited. A framework was either a base framework like TEI-C's TEI P5 framework, or it was based on a base framework. But since version 23.1+, the mechanism of *.framework files for configuring frameworks is replaced/supplemented with extension scripts. This allows us to design arbitrary tall stacks of frameworks, not only limited to height level 2. It's now possible to design base and intermediate frameworks with common functions. Only a thin layer is required for project-specific needs.

oXbytei¹ and oXbytao² are intermediate frameworks based on TEI P5. They are developed at the University of Münster and started as extractions of generic parts of a framework for the edition of the complete works of Ibn Nubata al Misri (1287-1366). They support the whole process from capturing and encoding up to making semantic annotations. Two design principles govern development:

1. Get as much configuration as possible from the TEI document's header! E. g. depending on the variant encoding declared in the header oXbytei produces a parallel segmentation, double end-point attached, or location referenced apparatus. Since not all information for setting up the editor is available in the header, oXbytei comes with its own XML configuration.
2. Introduce separation of concerns (SOC) and clean interfaces! The lack of SOC in Oxygen makes framework design hell for developers: E. g. you extend CSS with code for adding user dialogs and into this code, you again embed XPath expressions for accessing data and presenting selection options to the user. To overcome, oXbytei introduces interfaces for classes of operations, e. g. drawing a user dialog or accessing data for selection options. It then introduces loads of plugins that implement these interfaces, e. g. combo boxes, checkboxes, narrowing list selections as dialog plugins or XQuery and XSLT as data plugins. And finally, the XML config file binds the different plugin types to editing contexts. It can e. g. bind a multi-select check box dialog and an XQuery on your central witness catalog to the context of rdg/@wit.

Frameworks / Plugins that I learned from

- [ediarum.JAR](#)
- [BCDH TEI Completer](#)

Notes

1 <https://github.com/SCDH/oxbytei>,

2 <https://github.com/SCDH/oxbytao>

Poster: Automatic Validation, Packaging and Deployment of TEI Documents. What Continuous Integration can do for us

C. Lück

Universität Münster, Germany

Keywords: automation, validation, continuous integration, continuous deployment, error reports, quality control

Automatic Validation, Packaging and Deployment of TEI Documents

What Continuous Integration can do for us

Christian Lück

Keeping TEI documents under version control has many pros. Version control is not only a time machine that gives access to earlier versions of documents but distributed version control like Git makes a project robust against data losses and excels at merging the project members' work into a shared mainline. A feature of version control systems hardly known amongst textual scholars has become a key to professional software development: continuous integration (CI), i. e. automatic building and testing of what has been merged into the mainline. While TEI-C uses CI to automatically run tests on the guidelines and to generate schemas and views, CI has hardly been adopted by digital editions. The poster presentation wants to introduce it to a broader community and provides a template for automatic validation and deployment of TEI documents.¹

CI gets its strength from automation by running tests *regularly* and *uniformly*. For obvious reasons, CI has been transferred from software development to quality assurance of research data (in life sciences) by Cimiano et al. (2021). CI fires a pipeline of testing and deployment routines on certain events, e. g. a push to the mainline. Git servers, like Gitlab or Github, offer runtime environments for driving such pipelines.

However, the default output of a CI pipeline is a hurdle to widespread adoption of CI in scholarly editions, since it simply shows a log of the Linux console through a web interface. To break this barrier, the template's pipeline parses Jing's output and merges it with Schematron reports into a single human-readable report which is deployed on the git-servers publication environment (e.g. Gitlab pages). (Fig.) Thus, we get continuous validation and nice quality reports of our edition. On successful validation, a XAR package is assembled just as in a TEI publisher data template². Finally, it is deployed on a running eXist-db instance.

Summary

TEI documents tested: 3

Schematron-Reports: 3

Errors: **1** (Relax NG), **2** (Schematron)

Details

Document	Relax NG	Schematron
Heine-Traumbilder1.TEI-P5.xml	0	0
Trawr-Gesang.TEI-P5.xml	1	2
common.TEI-P5.xml	0	0

Technical Information:

Repository branch: <https://github.com/scdh/edition-data-template>

Repository slot name / base folder name: edition-data-template-cx

Fig.: Human readable quality report

References

Cimiano, Ph. et al. (2021): Studies in Analytic Reproducibility. The Conquaire Project. U Bielefeld Press. doi: 10.4119/unibi/2942780

1 <https://github.com/SCDH/edition-data-template-cx>.

2 <https://github.com/eeditiones/tei-publisher-data-template>

Poster: Adapting TEI for Braille

E. Forget

University of Toronto, Canada

Keywords: braille, bibliography, accessibility, book history, publishing

Adapting TEI for Braille

Ellen Forget

Bibliography as a field has undergone rapid changes to adapt for ever-evolving book formats in the digital age. Methods, tools, and techniques originally meant for manuscripts and printed books have now been adjusted to apply to ebooks, audiobooks, and other bookish objects. However, there is much less work currently available that considers the bibliographical differences of accessible book formats or, more specifically, braille as a book format. Braille lends itself well to analysis of materiality and format due to its tactile nature, but traditional bibliographical methods and tools were not developed with braille in mind and must be adapted to work with braille.

As part of a larger braille bibliography-focussed project, I am adapting TEI to work for analyzing braille books—specifically braille editions of illustrated children’s books. Illustrated books offer additional complexity to textual analysis that is compounded by the forced hierarchy of linear-text tools, and working with braille editions of illustrated books further complicates questions of hierarchy and format descriptions. How can textual scholars engage with text as data when there are multiple, overlapping (literally and metaphorically) sets of text in a single textual object? Can we extract a single data set from texts that communicate in multiple languages and codes, or with the addition of illustrations and images?

This poster will showcase the progress I have made so far in adapting TEI to work with braille, specifically using the multilingual prototype book as an example. The poster will touch on questions of textual hierarchy, line length/breaks, illustration descriptions, braille and format descriptions, and how languages are tagged, and it will include a wish list of TEI needs that I have not successfully adapted yet, as this is a work-in-progress project.

Keywords: braille, bibliography, accessibility, book history, publishing

Ellen Forget is a PhD student at University of Toronto in the Faculty of Information and the Book History and Print Culture program. They are also a graduate of Simon Fraser University’s Master of Publishing and Editing Certificate programs, and they work as a freelance editor. Their research interests include braille, accessible book formats, indie publishing, and speculative fiction genres.

Poster: Okinawan Lexicography in TEI: Challenges for Multiple Writing Systems

S. Miyagawa¹, K. Kato², M. Zlazli³, S. Machida⁴, S. Carlino⁵

1: National Institute for Japanese and Linguistics (NINJAL), Japan; 2: Tokyo University of Foreign Studies, Japan; 3: SOAS University of London, UK; 4: University of Hawai'i at Hilo, US; 5: Kyushu University/Hitotsubashi University, Japan

Keywords: lexicography, Okinawan, endangered language, multiple writing systems, language revitalization

Okinawan Lexicography in TEI: Challenges for Multiple Writing Systems So Miyagawa, Kanji Kato, Miho Zlazli, Seira Machida, and Salvatore Carlino

Okinawan is classified as one of the Northern Ryukyuan languages in the Japonic language family. It is primarily spoken in the south and central parts of the Okinawa Island of the Ryukyu Archipelago. It was the official lingua franca of the Ryukyu Kingdom and a literary vehicle, e.g., the *Omoro Soshi* poetry collection, but currently an endangered language. Okinawan has been recorded in various written forms: A combination of Kanji logograms and Hiragana syllabary with archaic spellings (e.g., *Omoro Soshi*) or modern spelling variations to approximate actual pronunciation, pure Katakana syllabary (e.g., Bettelheim's Bible translation), Latin alphabet (mostly by linguists), and pure Hiragana (popular).

The *Okinawago Jiten* (Okinawan Dictionary; OD), published by the National Institute for Japanese Language and Linguistics (NINJAL) in 1963 and revised in 2001^[1], uses the Latin alphabet for each lexical entry. We first added the possible writing forms listed above to the data in CSV format. We then converted the CSV into TEI XML using Python. Figure 1 presents a sample encoding of the TEI file for each entry. Here, we solved the multiple writing forms with <orth> tags with corresponding writing systems in @xml:lang attribute following BCP 47^[2] (e.g., xml:lang="ryu-Hira" for Okinawan words written in Hiragana). We added the International Phonetic Alphabet (IPA) and the accent type to make the pronunciation clearer with the <pron> tags.

```

<entry>
  <cit>
    <bibl> 『沖縄語辞典』 国立国語研究所資料集5第9刷 (2009), p.100</bibl>
  </cit>
  <form>
    <orth xml:lang="ryu-Hira">あびーぐいー</orth>
    <orth xml:lang="ryu-Jpan" n="1">叫び一声</orth>
    <orth xml:lang="ryu-Jpan" n="2">叫声</orth>
    <orth xml:lang="ryu-Latn" n="1">?abiigwii</orth>
    <orth xml:lang="ryu-Latn" n="2">abiigwii</orth>
    <pron notation="ipa">?abiigwii</pron>
    <pron notation="accent">0</pron>
  </form>
  <gramGrp>
    <pos>NOUN</pos>
    <subc>名</subc>
  </gramGrp>
  <sense xml:lang="jp-Jpan" n="1">
    <def>叫び声。kaamakara~nu cikariin. 遠くから叫び声が聞こえる。</def>
  </sense>
  <usg></usg>
</entry>

```

Fig. 1 TEI of each lexical entry

Using XSLT, we transformed this TEI file into a static webpage with a user-friendly GUI, as shown in Figure 2. It is anticipated that this digitization of OD and its publication under the open license will benefit key stakeholders, such as Okinawan heritage learners and worldwide Okinawan learners, being the largest Okinawan dictionary available online.

The screenshot shows a dictionary entry for '叫びゆん' (abi-yun). The page is annotated with blue boxes and lines identifying various linguistic features:

- PoS** (Part of Speech): 自動詞 (Automatic verb)
- Conjugation type**: =raN, =ti
- Alphabet**: ろーま字 (Romanized) *abiyun*, 国際音声字母 (International Phonetic Alphabet) *?abijun*, オリジナル (Original) *?abi=juN*
- Accent type**: ◎型
- Hiragana**: っあびゆん
- Kanji**: 叫びゆん
- Original**: ?abi=juN
- IPA**: ?abiigwii
- Sense 1**: 1. 叫ぶ。大声で呼ぶ。また、わめく。どなる。 (Call out loudly. Also, wameku. To be noisy.)
- Sense 2**: 2. 大声で泣く。泣き叫ぶ。 (Cry loudly. Cry and call out.)
- Sense 3**: 3. ほえる。犬・猫・豚などが鳴く。 (Bark. Dog, cat, pig, etc. barking.)

Fig. 2 Webpage rendition of TEI

Bibliography

^[1] National Institute for the Japanese Language and Linguistics (ed.), *Okinawago Jiten*, revised edition, Tokyo: Zaimusho Insatsukyoku, 2001.

^[2] Text Encoding Initiative, “teidata.language,” P5: Guidelines for Electronic Text Encoding and Interchange, version 4.4.0, last updated on 19th April 2022, revision ff9cc28b0, <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-teidata.language.html> (last accessed on 20th June 2022).

Biographies

So Miyagawa is an assistant professor at the National Institute for Japanese Language and Linguistics (NINJAL), Tokyo, Japan. His research interests are in digitizing rare documents and audios of endangered languages and dialects. Currently, he is working on digital corpora and dictionaries of Ryukyuan languages using TEI as part of the NINJAL Endangered Language Digital Archive & Library (NELDAL) project. He obtained Dr.phil. from the University of Göttingen, Germany.

Kanji Kato is a Ph.D. student at the Tokyo University of Foreign Studies. He is working on documentation of the Tokunoshima dialect of the Amami language, one of the Northern Ryukyuan languages. He is also interested in the digitization of various language materials. He is also a programmer of the NINJAL Endangered Language Digital Archive & Library (NELDAL) project.

Miho Zlazli is an Okinawan doctoral student at the SOAS University of London, the United Kingdom. Her research interests include new speakers of minoritized languages, indigenous research paradigm, and language documentation. She is a language revitalization advisor of the NINJAL Endangered Language Digital Archive & Library (NELDAL) project.

Seira Machida is a Ph.D. candidate at the University of Hawai'i at Hilo. She is writing her doctoral dissertation on language revitalization, focusing on the Okinawan language. At the same time, she is currently doing fieldwork on Okinawa Island. She is a lexicographer of the NINJAL Endangered Language Digital Archive & Library (NELDAL) project.

Salvatore Carlino is an adjunct researcher at Kyushu University (Fukuoka, Japan) and Hitotsubashi University (Tokyo, Japan). His main interests are descriptive linguistics and documentation of the Ryukyuan languages. He published the first descriptive grammar of the Iheya dialect of Okinawan as his doctoral thesis. He has also created the Online Dictionary of the Japonic Languages as a post-doc project, an online dictionary for the Japonic languages. He is now a part-time worker of the NELDAL.

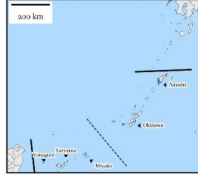
Okinawan Lexicography in TEI: Challenges for Multiple Writing Systems

1 So Miyagawa, 2 Kanji Kato, 3 Miho Zlazli, 4 Seira Machida, and 5 Salvatore Carlino

1: National Institute for Japanese and Linguistics (NINJAL), Japan; 2: Tokyo University of Foreign Studies/JSFS/NINJAL, Japan; 3: SOAS University of London, UK; 4: University of Hawai'i at Hilo, US; 5: Kyushu University/Htotsubashi University, Japan

Okinawan and Multiple Spelling Problems

- Okinawan
 - Classified as one of the Northern Ryukyuan languages in the Japonic language family
 - Primarily spoken in the south and central parts of the Okinawa Island
 - Official lingua franca of the Ryukyu Kingdom and a literary vehicle, e.g., the Omoro Soshi poetry collection
 - Currently an endangered language
- Various written systems used
 - A combination of Kanji logograms and Hiragana syllabary with archaic spellings
 - Modern spelling variations to approximate actual pronunciation
 - Pure Katakana syllabary (e.g., Bettelheim's Bible translation)
 - Latin alphabet (mostly by linguists)
 - Pure Hiragana (popular)



```
<entry>
  <cit>
    <bibl>『沖縄語辞典』国立国語研究所資料集5第9刷 (2009), p.100</bibl>
  </cit>
  <form>
    <orth xml:lang="ryu-Hira">あびー</orth>
    <orth xml:lang="ryu-Jpan">あびー</orth>
    <orth xml:lang="ryu-Jpan">あびー</orth>
    <orth xml:lang="ryu-Latn">abi</orth>
    <orth xml:lang="ryu-Latn">abi</orth>
    <pron notation="ipa">ʔabiŋwi</pron>
    <pron notation="accent">0</pron>
  </form>
  <gramGrp>
    <pos>NOUN</pos>
    <sub>名</sub>
  </gramGrp>
  <sense xml:lang="jp-Jpan" n="1">
    <def>あびー声。kaamakara~nu cikariin. 遠くからあびー声が聞こえる。</def>
  </sense>
  <usg></usg>
</entry>
```

Digitization of NINJAL's Okinawan Dictionary

- The *Okinawgo Jiten* (Okinawan Dictionary; OD): published by the National Institute for Japanese Language and Linguistics (NINJAL) in 1963 and revised in 2001; the Latin alphabet for each lexeme
- Addition of the possible writing forms listed above to the data in CSV format
- Conversion of the CSV into TEI XML using Python
- Solution of the multiple writing forms with <orth> tags with corresponding writing systems in @xml:lang attribute following BCP 47
 - E.g., xml:lang="ryu-Hira" for Okinawan words written in Hiragana
- International Phonetic Alphabet (IPA) and the accent type to make the pronunciation clearer with <pron> tags

A	B	C	D	E	F	G	H	I	J
辞書	発出形態	アクセント	品詞	文類などの種類	補綴	意味 1	意味 2	意味 3	意味 4
1	あびー								
51	Tabuucan	(1)	自	自	自	あびー	あびー		
52	Tabui	(0)	名			あぶみ			
53	Tabuiku	(0)	名			あぶみ			
54	Tabuigu	(0)	名			あぶみ			
55	Tabuusa	(1)	名	文	文	あぶみ			
56	Tabusi	(0)	名			あぶみ			
57	Tabusibene	(0)	名			あぶみ			
58	Tabusimaku	(0)	名	文	文	あぶみ			
59	Tabusimici	(0)	名			あぶみ			
60	Taca	(0)	名			あぶみ			
61	Taca?aa	(0)	名			あぶみ			
62	Tacaga	(1)	名			あぶみ			
63	Tacaga?in	(1)	自	自	自	あぶみ			
64	Tacajuu	(0)	名			あぶみ			
65	Tacajusa?in	(1)	自	自	自	あぶみ			
66	Tacajusumisi	(0)	名			あぶみ			
67	Taci	(1)	名	文	文	あぶみ			
68	Tacibee	(0)	名			あぶみ			
69	Tacibi	(0)	名			あぶみ			
70	Tacibiraci	(0)	名			あぶみ			
71	Taciguru	(0)	名			あぶみ			
72	Tacihianstu	(0)	名			あぶみ			

User-friendly dictionary website

Using XSLT, we transformed this TEI file into a static webpage with a user-friendly GUI, as shown in the figure below. It is anticipated that this digitization of OD and its publication under the open license will benefit key stakeholders, such as Okinawan heritage learners and worldwide Okinawan learners, being the largest Okinawan dictionary available online.



This research is supported by National Institutes for the Humanities' project "Building a Digital Library of Academic Knowledge."

Poster: Text as Object: Encoding the data for 3D annotation in TEI

J. Ogawa¹, K. Nagasaki², I. Ohmukai³, Y. Nakamura³, A. Kitamoto¹

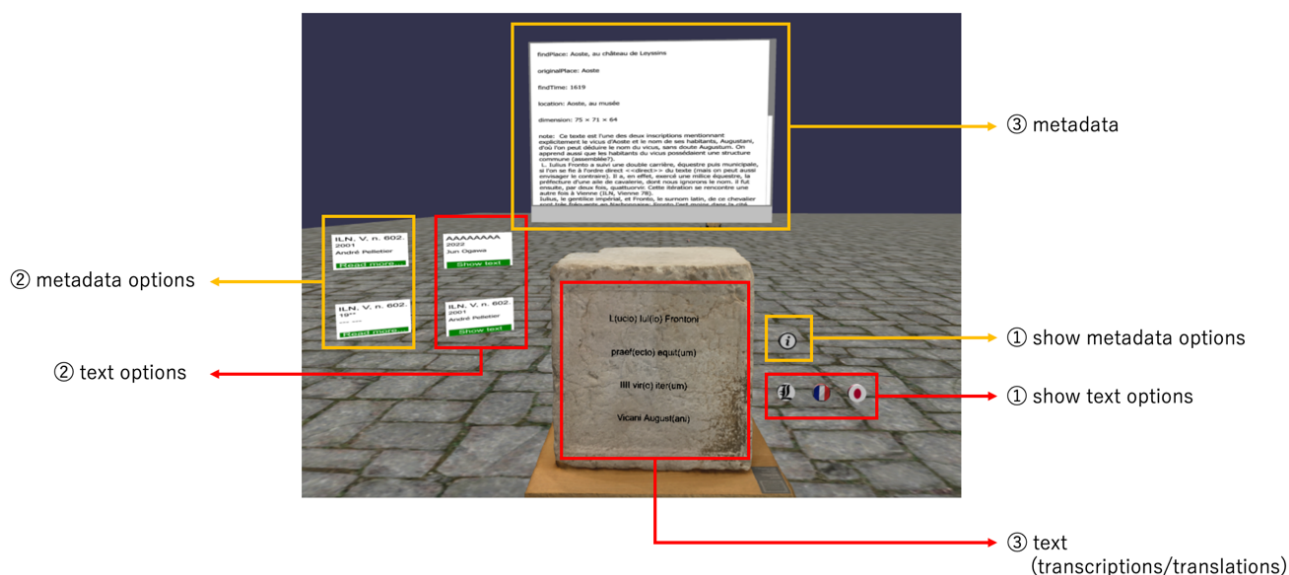
1: Center for Open Data in the Humanities, Japan; 2: International Institute for Digital Humanities, Japan; 3: University of Tokyo, Graduate School of Humanities and Sociology

Keywords: 3D scholarly editions, annotation, <sourceDoc>, Babylon.js

Text as Object: Encoding the data for 3D annotation in TEI

Jun Ogawa, Kiyonori Nagasaki, Ikki Ohmukai, Yusuke Nakamura, Asanobu Kitamoto

Recently, the concept of 3D scholarly edition or 3D documentation has been discussed in the field of Digital Humanities [1][2]. In this study, we have encoded several pieces of ancient epigraphy in TEI and attempted to create a 3D edition by putting scholarly annotations such as metadata, transcriptions, translations, and commentaries. Our implementation enables users to analyze the materials interactively and virtually by exploring the related information in the 3D edition.



For this edition, we need to consider three types of data: 1) texts and apparatus, 2) disposition of texts on the object, and 3) construction of GUIs. It would be reasonable to utilize the TEI guideline for 1) and 2) since they deal with textual issues like representation of different readings and interpretation, or spatial organization of textual elements.

We currently think of using <sourceDoc>, an element prepared for encoding the diplomatic edition of the 2D text, to represent a 3D object with @rend="3D".


```

<sourceDoc rend="3D">
  <surfaceGrp type="3DObject">
    <figure><graphic url="http://3Dmodel.glb"></graphic></figure>
    <surface type="textPosition"></surface>
    <surface type="text" xml:lang="la" ana="original"></surface>
    <surface type="text" xml:lang="ja" ana="translation"></surface>
  </surfaceGrp>
</sourceDoc>

```

To identify the position of the texts, we use <zone> in <surface> as follows.

```

<surface xml:id="sf_00" type="textPosition">
  <zone xml:id="z_01_sf_00" ulx="-0.02" uly="5.51" ana="z:-0.27"/>
  <zone xml:id="z_02_sf_00" ulx="-0.02" uly="5.41" ana="z:-0.27"/>
  <zone xml:id="z_03_sf_00" ulx="-0.02" uly="5.31" ana="z:-0.27"/>
  <zone xml:id="z_04_sf_00" ulx="-0.02" uly="5.21" ana="z:-0.27"/>
</surface>

```

We put other <surface> to represent textual contents including apparatus.

```

<surface xml:id="sf_01" type="text" ana="original" xml:lang="lat" source="#wit_1 #wit_2">
  <line xml:id="z_01_sf_01" corresp="#z_01_sf_00">
    L(ucio) <app><lem wit="#wit_1">Iul(io)</lem><rdg wit="#wit_2">Iul(ius)</rdg></app> Frontoni
  </line>
  <line xml:id="z_02_sf_01" corresp="#z_02_sf_00">
    praef(ecto) equit(um)
  </line>
  <line xml:id="z_03_sf_01" corresp="#z_03_sf_00">
    IIII <app><lem wit="#wit_1">vir(o)</lem><rdg wit="#wit_2">vir(orum)</rdg></app> iter(um)
  </line>
  <line xml:id="z_04_sf_01" corresp="#z_04_sf_00">
    Vicani August(ani)
  </line>
</surface>

```

While we, for now, describe the plain text according to the Leiden convention, we plan to apply *Epidoc* markup in our edition. The data other than the text itself, such as witnesses, metadata, and commentaries are stored in <listWit> under <standOff>.

Even if our study is still in the proposal stage, it paves the way for creating the 3D edition with TEI. Considering that TEI has not yet provided an established way of encoding texts on 3D materials, this study might serve as a springboard for a wider discussion on the features of ‘text’ in 3D virtual space and their practical implementation in the TEI framework, such as the redefinition of <sourceDoc> or the coordinate description in 3D.

- [1] Schreibman, S., Papadopoulos, C. (2019), ‘Textuality in 3D: three-dimensional (re)constructions as digital scholarly editions’, *International Journal of Digital Humanities*, vol. 1, pp. 221-233.
- [2] Vitale, V. (2017), *Rethinking 3D Visualisation: From photorealistic visual aid to multivocal environment to study and communicate cultural heritage* [Doctoral Dissertation, King’s College London], King’s Research Portal:

Poster: Explainable Supervised Models for Bias Mitigation in Hate Speech Detection: African American English

A. Gabriel, M. Sinclair

Northumbria University

Keywords: Natural Language Processing, Explainable AI, Computing, Social Media, Hate Speech

Explainable Supervised Models for Bias Mitigation in Hate Speech Detection: African American English.

**Aaron Gabriel and Dr Mark Sinclair
Northumbria University**

Abstract

Automated hate speech detection systems have great potential in the realm of social media but have seen their success limited in practice due to their unreliability and inexplicability. Two major obstacles they have yet to overcome is their tendency to underperform when faced with non-standard forms of English and a general lack of transparency in their decision-making process. These issues result in users of low-resource languages (those that have limited data available for training) such as African-American English being flagged for hate speech at a higher rate than users of mainstream English. The cause of the performance disparity in these systems has been traced to multiple issues including social biases held by the human annotators employed to label training data, training data class imbalances caused by insufficient instances of low-resource language text and a lack of sensitivity of machine learning (ML) models to contextual nuances between dialects. All these issues are further compounded by the 'black-box' nature of the complex deep learning models used in these systems. This research proposes to consolidate seemingly unrelated recently developed methods in machine learning to resolve the issue of bias and lack of transparency in automated hate speech detection. The research will utilize synthetic text generation to produce a theoretically unlimited amount of low-resource language text training data, machine translation to overcome annotation conflicts caused by contextual nuances between dialects and explainable ML (including integrated gradients and instance-level explanation by simplification). We will attempt to show that when repurposed and integrated into a single system these methods can both significantly reduce bias in hate speech detection tasks whilst also providing interpretable explanations of the system's decision-making process.

Poster: A TEI/IIIF Structure for Adding Palaeographic Examples to Catalogue Entries

S. M. Winslow

University of Graz, Austria

Keywords: manuscript studies, palaeography, IIIF, cataloguing

A TEI/IIIF Structure for Adding Palaeographic Examples to Catalogue Entries

S. M. Winslow

University of Graz, Austria

The study of palaeography generally relies on either expert testimony with sparse examples or separate, specialist catalogues imaging and documenting the specific characteristics of each hand. Both practices presumably made much more sense due to the cost, difficulty, and space used by printed catalogues in the past, but with modern practice in cataloguing manuscripts via TEI and disseminating images via IIIF, these difficulties have been largely obviated. Accordingly, it is desirable to have a simple, consistent, and searchable way to embed examples of manuscript hands within the TEI, as a companion to elements from msdescription that describe hand features. This poster will demonstrate a simple and re-useable structure for embedding information about the palaeography of manuscript hands in msdescription and associating it with character examples using IIIF. An example implementation, part of the Hidden Treasures from the Syriac Manuscript Heritage project, will be demonstrated and an ODD containing the new elements and structure will be made available.

A TEI/IIIF Structure for Standardizing Palaeographic Examples

Sean M. Winslow



Summary

The wealth of palaeographical examples described in catalogue entries that are stored in TEI are not currently comparable or retrievable in any meaningful way. A standard for the markup of palaeographic examples with corresponding IIIF zones allowing the serving of the associated images would allow for the easy indexing and comparison of hands. Another advantage of linking to small regions in the IIIF is that the context is preserved for scholars who follow the link back to the original.

A simple standard, developed for a project on Syrian manuscripts, but with general utility, is presented here and described in the linked ODD. It provides three new elements within `handDesc/handNote` to link description to examples: `letterformDesc`, `letterform`, and `letterformExample`, intended to promote discoverability and reusability.

Markup example:

```
<handDesc>
  <handNote scope="major" script="syr"
  Syrc" xml:id="handNote1">
    <letterformDesc>
      <letterform></letterform>
      <letterformExample>
        <graphic url="...
        /pct:27.13816,64.01681,2.38487,1.55336/full/
        0/default.jpg"/>
      </letterformExample>
      <letterformExample>
        <graphic url="...
        /pct:41.99561,66.10015,2.38487,1.55336/full/
        0/default.jpg"/>
      </letterformExample>
      <letterformExample>
        <graphic url="...
        /pct:36.48465,31.12208,2.46711,1.97368/full/
        0/default.jpg"/>
      </letterformExample>
    </letterformDesc>
    <note>
      the right side of the Heh is not
      curving but bit straight
    </note>
  </letterformDesc>
</handDesc>
```

Letter	Form	Example	Notes
ⲙ			the right side of the Heh is not curving but bit straight
ⲁ			
ⲛ			
ⲛ			
ⲛ			
ⲛ			
ⲛ			
ⲛ			

The first hand used a brown ink, now paled. It had been carefully overwritten at many places in the book applying a dark brown ink (see the contrast e.g. on fol. 106. Headlines written in red. All texts (old and new)

Fig. 1: Example of proposed practice: Damascus, Syriac Orthodox Patriarchate Library, SOP 92 Anonymous. Cause of Causes WIP preview of the Hidden Treasures from the Syrian Heritage interface (PI: Erich Renhart)

Find the ODD on GitHub

https://github.com/larkvip/PalaeographyExample_ODD

Let's automate output like this across projects:

Fig. 2: Output from the DASH project (no relation to current project, except as inspiration) dash.stanford.edu/viewer/

Acknowledgements

- The work in this poster was funded by the FWF-Project "Syriac Manuscript Treasures" (PI: Erich Renhart)
- Markup: Ephrem Aboud Ishaac, Roger Akhrass, and Erich Renhart
- Feedback and proofreading: Elisabeth Steiner

Contact information

ZENTRUM FÜR INFORMATIONSMODELLIERUNG
AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

- Email: sean.winslow@uni-graz.at
- <https://informationsmodellierung.uni-graz.at>

(This poster CC BY-NC 4.0)

Poster: From facsimile to online representation. The Centre for Digital Editions in Darmstadt. An Introduction

K. Fischer, S. Kalmer, D. Kampkaspar, S. Müller, M. Scheffer, M. E.-H. Seltmann, K. Wunsch
University and State Library Darmstadt, Germany

Keywords: Digital edition, projects, cooperation, digital texts, infrastructure

From facsimile to online representation. The Centre for Digital Editions in Darmstadt. An Introduction

K. Fischer, S. Kalmer, D. Kampkaspar, S. Müller, M. Scheffer, M. E.-H. Seltmann, K. Wunsch

University and State Library Darmstadt, Germany; katrin.fischer@tu-darmstadt.de,
silke.kalmer@tu-darmstadt.de, dario.kampkaspar@tu-darmstadt.de,
tonia-sophie.mueller@tu-darmstadt.de, marc.scheffer@tu-darmstadt.de,
melanie.seltmann@tu-darmstadt.de, kevin.wunsch@tu-darmstadt.de

The Centre for Digital Editions in Darmstadt (CEiD) covers all aspects of preparing texts for digital scholarly editions from planning to publication. It not only processes the library's own holdings, but also partners with external institutions.

Workflow

After applying both automatic and manual methods for text recognition (OCR/HTR) the output is used as a starting point for the realisation of the digital edition as an online publication. In addition, a variety of transformation tools is used to convert texts from different formats such as XML, JSON, WORD-DOCX or PDF into TEI-based formats (TEI Consortium 2022), thus substantially enabling uniformity across different projects. These texts can be annotated and enriched with metadata. Furthermore, entities can be marked up, which are managed in a central index file. This workflow is not static, but can be adapted according to the needs of the project. Scholars and developers alike can benefit from this workflow which centers on translating various data formats into TEI.

Framework

The XML files are stored in eXist-db (eXist Solutions 2022) and presented in various user-friendly ways with the help of the framework wdbplus (Kampkaspar 2018), which is designed according to the needs of large institutions with diverse corpora, such as a university library. By default, the transcribed text and the corresponding scan presented side by side. Additionally, different forms of presentation are available so that the special needs of individual projects can be considered. Further advantages of wdbplus are various REST-APIs, which not only allow the retrieval of individual texts, but also of metadata and further information. Full-text search is realised at project level as well as across projects. CEiD's portfolio includes several projects in which a multitude of texts are processed. The source material ranges from early modern prints and manuscripts to more recent texts and

includes early constitutional texts, religious peace agreements, newspapers and handwritten love letters.

Bibliography

- **eXist Solutions** (2022): *eXist-db* [Online]. Available at: <https://exist-db.org> (Accessed: 20 June 2022)
- **Kampkaspar**, Dario (2018): *W. Digitale Bibliothek (wdbplus)*. Available at: <https://github.com/dariok/wdbplus> (Accessed: 20 June 2022)
- **TEI Consortium** (2022) *TEI P5: Guidelines of Electronic Text Encoding and Interchange*. [Version 4.4.0]. Available at: <https://tei-c.org/guidelines/p5/> (Accessed: 20 June 2022)

Biographies

Katrin Fischer currently works on a project about letters in the 18th Century at the Centre for Digital Editions (CEiD). She holds a degree in History and Applied Geosciences. Her research interests include indexes in scholarly editions and mapping historical networks.

Silke Kalmer is a staff member at the Centre for Digital Editions in Darmstadt (CEiD) at the University and State Library Darmstadt. She holds a master of arts degree in Linguistic and Literary Computing. At CEiD she is involved in various projects surrounding digital editions. Her interests lie in implementing methods of the digital humanities for using and creating digital editions.

Dario Kampkaspar is the head of the Centre for Digital Editions in Darmstadt. He holds a *Magister Artium* in history and English philology and has been working on TEI-based projects at the Duke Augustus Library in Wolfenbüttel, the Austrian Academy of Sciences in Vienna and the CEiD for more than 10 years.

Marc Scheffer is a project staff at the Centre for Digital Editions in Darmstadt (CEiD). He holds a degree in Japanese studies and currently works on a project on cooperation in Open Access publishing. His research interests include multilingual digital scholarly editions.

Melanie Seltmann currently works at the Centre for Digital Editions (CEiD) at the University and State Library Darmstadt in the project *Gruß & Kuss* as well as in the NFDI Constortium *Text+*, Task Area Editions. She is interested in Digital Humanities, Citizen Science, corpus linguistics, variational linguistics, standards & best practices, and annotation. In her PhD thesis she is focusing on annotations.

Kevin Wunsch currently works at the Centre for Digital Editions at the University and State Library Darmstadt. He is the Digital Humanist in the project “European Peace Agreements Digitally”. Besides that, he is a PhD-Student at Heidelberg University, focusing on George Anson’s circumnavigation with digital methods, such as Topic Modelling and digital editing.

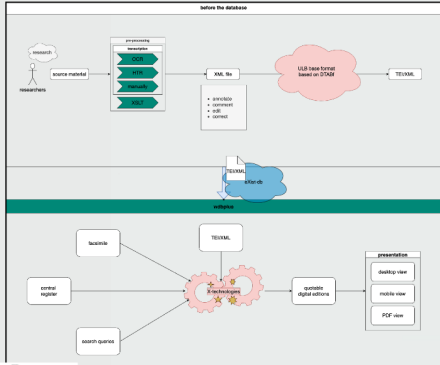
~10 employees

10 centuries covered

60k+ texts

12 ongoing projects

4 servers



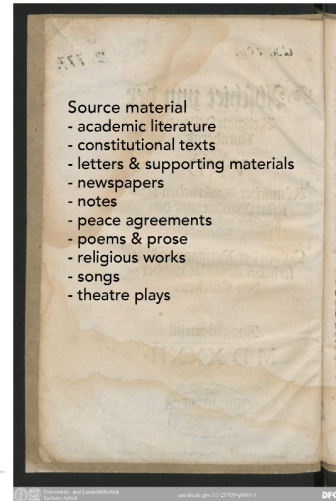
Third-party funded projects

- <The> Book of Letters of [Hildegard of Bingen](#)
- Digitisation of the [Darmstädter Tagblatt \(1740 – 1986\)](#)
- [European Religious Peace Agreements – a Digital Edition](#)
- <The> German Letter in the 18th Century
- [Greetings and kisses – letters digital. Citizens receive love letters](#)
- [Indexing and Digitisation of the Baron von Hüpsch Collection](#)
- [OATbyCO – Open Access Transformation by Cooperation](#)
- [Text+](#)

Further cooperations

- Briefnetzwerke der Brüder Grimm
- Early Modern German Texts
- Johann Arndts (1555-1621) Vier Bücher von wahrem Christentum (1610). Kritische, kommentierte Hybrid-Edition.
- [IEG: Controversia et Confessio](#)
- IEG retro

Further projects planned.



- Source material
- academic literature
 - constitutional texts
 - letters & supporting materials
 - newspapers
 - notes
 - peace agreements
 - poems & prose
 - religious works
 - songs
 - theatre plays

Poster: From Oxgarage to TEIGarage and MEIGarage

P. Stadler, A. Ferger, D. Röwenstrunk

Paderborn University, Germany

Keywords: Software Sustainability, Software Development, DH Communities

From Oxgarage to TEIGarage and MEIGarage

The OxGarage is a „a web, and RESTful, service to manage the transformation of documents between a variety of formats“¹. It was originally developed by Poznan Supercomputing and Networking Center and Oxford University Computing Services for the EU-funded ENRICH project (Cummings et al 2009) and served the DH community for more than a decade.² In 2019 the OxGarage was chosen as a fundament for creating an MEIGarage, a “workshop’ for symbolic music encoding data”. Alike the OxGarage, the MEIGarage serves the MEI community as a web and RESTful service to facilitate transformations and interoperability of their specific file formats, e.g. MEI, MusicXML, and Lilypond, but also from symbolic music notation to audio file formats.

With the formation of the German NFDI and its NFDI4Culture consortium³, a dedicated workload was granted for further development on MEIGarage. Although this grant is offered to improve the music related features, the common code base of OxGarage and MEIGarage should stay strong and be developed further, for the benefit of both communities. As a first step, the repository structure was reorganized while keeping the Git history, resulting in various new repositories, so that each module/part has its own development space, version history and numbering, releases, etc. All the configuration is left to a wrapper repository that pulls together the parts required for this particular “X”Garage.

Hence, a new name for the TEI OxGarage was needed: both to emphasize its specific use for the TEI community and to strip off the now outdated Oxford legacy.

Besides restructuring and refactoring the code base, there’s some improvements and new features to be announced:

- building of the libraries and Docker images automated further,
- security issues concerning log4j fixed,
- Tomcat, Java, and other dependencies updated,
- validation functionality extended for TEI files,
- Swagger OpenApi documentation added.

For the software to become more sustainable and up-to-date badges⁴ are currently evaluated and will be used in future releases.

Bibliography

- James Cummings, Tomasz Parkoła, Mariusz Stanisławczyk and Marcin Werla: Report on the Documentation and Use of the ENRICH Garage Engine (ENRICH Reports on Deliverables WP3 D 3.4). 30 October 2009. Online at http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/ENRICH_WP3_D3.4pt1_EGE_0_0.pdf

Notes

1 <https://github.com/TEIC/oxgarage#readme>

2 Projects using the OxGarage include e.g. DHConvalidator, Roma, RomaBeta, jewish-history-online.net, and Music Performance Markup.

3 <https://nfdi4culture.de>

4 <https://fair-software.eu/> and <https://bestpractices.coreinfrastructure.org>

From OxGarage to TEIGarage and MEIGarage

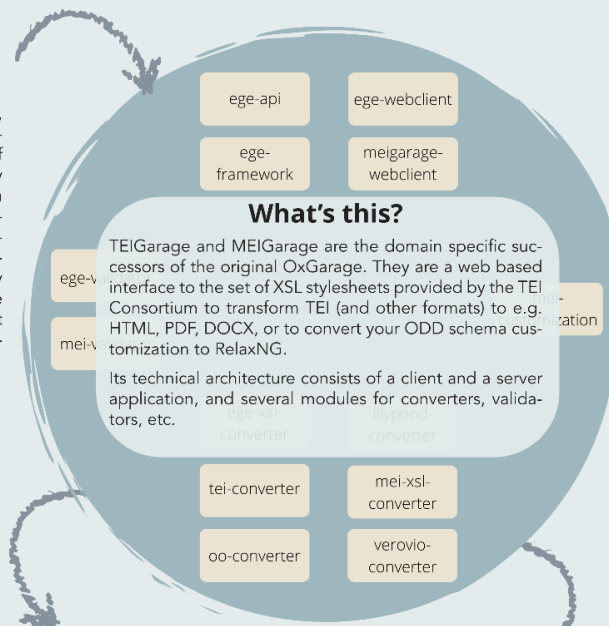
Peter Stadler (Paderborn University)
Anne Ferger (Paderborn University)
Daniel Rößenstrunk (Paderborn University)



10.5281/zenodo.7061525

Legacy OxGarage

The OxGarage is a „a web, and RESTful, service to manage the transformation of documents between a variety of formats“. It is written in Java and was originally developed by Poznan Supercomputing and Networking Center and Oxford University Computing Services for the EU-funded ENRICH project and served the DH community for more than a decade.



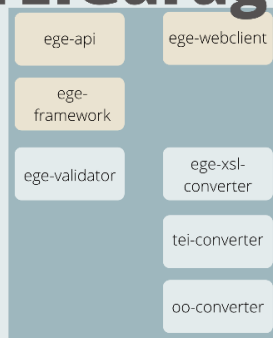
Refactoring

For facilitating the maintenance and development of "XGarages", the original OxGarage Git repository was split along the modular code structure (while keeping the Git history).

Every module now has its own development space, version history and numbering, releases, etc.

Java version and various dependencies have been upgraded, and security issues have been fixed.

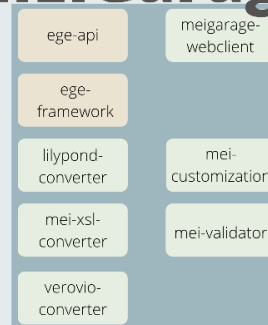
TEIGarage



New Features

- Validation functionality extended for TEI files
- Latest release of TEI Guidelines, sources, and stylesheets included in the Docker image
- Tomcat upgrade to version 9
- Swagger OpenAPI

MEIGarage



Goals

- Improved software sustainability by adhering to FAIR standards and OpenSSF Best Practices.
- Archive the OxGarage repo
- Switch from oxgarage.tei-c.org to teigarage.tei-c.org

Poster: Towards a digital documentary edition of CCCC41: The TEI and Marginalia-Bearing Manuscripts

P. O Connor

University of Oxford, United Kingdom

Keywords: marginalia, Old English, mise-en-page, sourceDoc, facsimile

Towards a digital documentary edition of CCCC41: The TEI and Marginalia-Bearing Manuscripts

Patricia O Connor, University of Oxford

The specific aim of this case study is to demonstrate how the TEI Guidelines have transformed the representation of an important corollary of the medieval production process; the annotations, glosses and other textual evidence of an interactive engagement with the text. Cambridge, Corpus Christi College MS 41 (CCCC MS 41) best exemplifies the value of the TEI in this respect as this manuscript is noted for containing a remarkable record of textual engagement from early medieval England. CCCC MS 41 is an early-eleventh century manuscript witness of the vernacular translation of Bede's *Historia ecclesiastica*, commonly referred to as the *Old English Bede*. However, in addition to preserving the earliest historical account of early medieval England, the margins of CCCC MS 41 contain numerous Old English and Latin texts. Of the 490 pages of CCCC MS 41, 108 pages contain marginal texts which span several genres of Old English and Latin literature; and thereby provide the potential for substantial evidence of interaction with the manuscript's central text.

While the marginalia of CCCC MS 41 continue to excite scholarly attention, the representation of this vast body of textual engagement poses certain challenges to editors of print scholarly editions. This poster emphasises the importance of the transcription process in successfully conveying the *mise-en-page* of marginalia-bearing manuscripts and explains how adopting the <facsimile> or <sourceDoc> approach encourages further engagement with and a deeper understanding of CCCC MS 41's marginalia.

This discussion of CCCC MS 41 and the representation of marginalia endeavours to demonstrate how transcribing marginalia-bearing manuscripts using a non-interpretative approach has the potential to significantly enlighten our understanding of medieval reading and scribal practices in early medieval England, as well as challenge our preconceptions about marginalised texts.

Poster: Transatlantic Networks - a Pilot: mapping the correspondence of David Bailie Warden (1772-1845)

J. Orr, S. Howard, J. Cummings

Newcastle University, United Kingdom

Keywords: letters, America, France, transnational, networks

Transatlantic Networks - a Pilot: mapping the correspondence of David Bailie Warden (1772-1845)

J. Orr, S. Howard, J. Cummings

Newcastle University, United Kingdom

The scientific revolution of the nineteenth century is often seen as remediating the early modern republic of letters (Klancher) from the pens of learned individuals to learned Institutions. This project aims to map the transatlantic network of one of the most important hubs in the exchange of literary and scientific correspondence, David Bailie Warden (1772-1845). Warden is known as an Irish political asylum seeker, American diplomat, and respected Parisian scientific writer in his own right, authoring and collaborating in foundational statistical works on America, the burgeoning natural sciences, and anti-slavery. More importantly, his correspondence with at least 3000 individuals and learned institutions reframes our perspective on the scientific revolution, its historical context, and its everyday activities. In addition to traditional close reading methods, this project tests methods from the field of scientific network analysis to enable us to identify other important network nodes, enabling the process of continual discovery. This project seeks to compile not only a 'who's who' of the intellectual community in this period but to identify previously hidden facilitative figures whose importance to the fabric of the republic of letters might not be at first obvious due to a range of marginalising factors including: social class, transnationality, gender, religion, or other liminal identities.

Conference Sessions – Thursday 15 September 2022

Session 4A – Short Papers – 09:30 - 11:00

Session 4A: Short-Papers

Location: ARMB: 2.98

Chair: Peter Stadler, Paderborn University

Short Paper: TEI and the Re-Encoding of Born-Digital and Multi-Format Texts

E. Forget, A. Galey

University of Toronto, Canada

Keywords: digital texts, textual studies, born-digital, electronic literature

TEI and the Re-Encoding of Born-Digital and Multi-Format Texts

Ellen Forget and Alan Galey

What affordances can TEI encoding offer scholars who work with born-digital, multi-format, and other kinds of texts produced in today's publishing environments, where the term "digitization" is almost redundant? How can we use TEI and other digitization tools to analyze materials that are already digital? How do we distinguish between a digital text's multiple editions or formats and its paratexts, and what differences do born-digital texts make to our understanding of markup? Can TEI help with a situation such as the demise of Flash, where the deprecation of a format has left many works of electronic literature newly vulnerable—and, consequently, newly visible as historical artifacts?

These questions take us beyond descriptive metadata and back to digital markup's origins in electronic typesetting, but also point us toward recent work on electronic literature, digital ephemera, and the textual artifacts of the very recent past. Drawing from textual studies, publishing studies, book history, disability studies, and game studies, we are experimenting with the re-encoding of born-digital materials, using TEI to encode details of the texts' form and function as digital media objects. In one case, we are working with an exclusively born-digital source—the Flash-based website *ApertureScience.com*, which was a paratext of the video game *Portal*—but our main example was published simultaneously in analogue and digital formats: the born-accessible book *Disfigured* by Amanda Leduc, which was published in seven formats, including braille and other accessible formats.

Drawing on our initial encoding and modelling experiments, this paper explores the affordances of using TEI and modelling for born-digital and multi-format textual objects, particularly emerging digital book formats.

Ellen Forget is a PhD student at University of Toronto in the Faculty of Information and the Book History and Print Culture program. They are also a graduate of Simon Fraser University's Master of Publishing and Editing Certificate programs, and they work as a freelance editor. Their research interests include braille, accessible book formats, indie publishing, and speculative fiction genres.

Alan Galey is Associate Professor in the Faculty of Information at the University of Toronto, cross-appointed to the Department of English. He is the author of *The Shakespearean Archive: Experiments in New Media from the Renaissance to Postmodernity* (Cambridge, 2014), as well as several journal articles and book chapters. His current research focuses on methods for the bibliographical study of digital texts and artifacts, from ebooks to videogames to digital recordings of musical performances (veilofcode.wordpress.com).

Short Paper: Capturing the Thread Structure: A Modification of CMC-Core to Account for Characteristics of Online Forums

S. Reimann, L. Rodenhausen, F. Elwert, T. Scheffler
Ruhr-University Bochum, Germany

Keywords: online forum, thread structure, social media, computer mediated communication

Capturing the Thread Structure: A Modification of CMC-Core to Account for Characteristics of Online Forums

Sebastian Reimann, Lina Rodenhausen, Frederik Elwert, Tatjana Scheffler (Ruhr University Bochum, Germany)

Representing computer mediated communication (CMC), such as discussions in online forums, according to the guidelines of the Text Encoding Initiative was addressed by the CMC Special Interest Group whose proposal was recently integrated in the TEI guidelines. However, these guidelines have a general aim and are not specifically tailored to capturing the thread structure of online forums.

Not only do online forums as a whole differ from other forms of CMC, but there are often also considerable differences between individual platforms. We created a corpus of posts from various religious online forums, including different communities on Reddit, as well as two German forums which specifically focus on the topic of religion, with the purpose of analyzing their structure and textual content. These forums differ in the way threads are structured, how emojis are used, and how people are able to react to other posts, for example by voting (see Figs. 1,2).

We present a twofold contribution to represent the features of online forums as a genre and still be flexible enough to enable the encoding of a wide range of different online forums. On the one hand, we slightly modify the attributes of the add and person elements to better account for edits to posts and deleted users. Additionally, we argue that elements from the gajim module may be a good fit for representing non-unicode emojis. On the other hand, we demonstrate ways to apply the general TEI guidelines to a specific form of CMC. In our data, thread structures with nested replies are of central importance. We therefore propose to represent nested threads using hierarchical xml structures (Fig. 3) instead of the artificially flattened structure using @indentLevel proposed in the TEI-CMC guidelines. This solution also stays closer to the familiar view on forum websites.

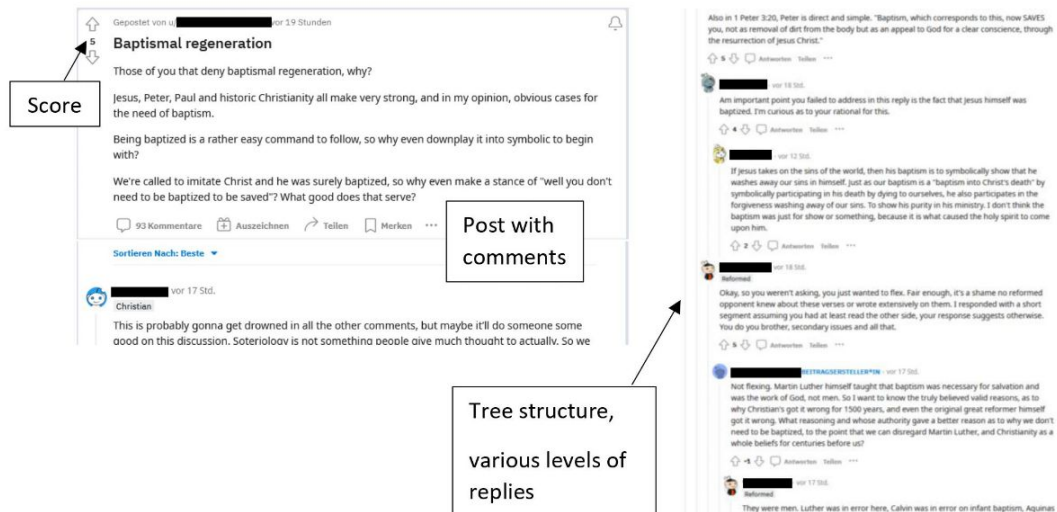


Fig. 1: Reddit

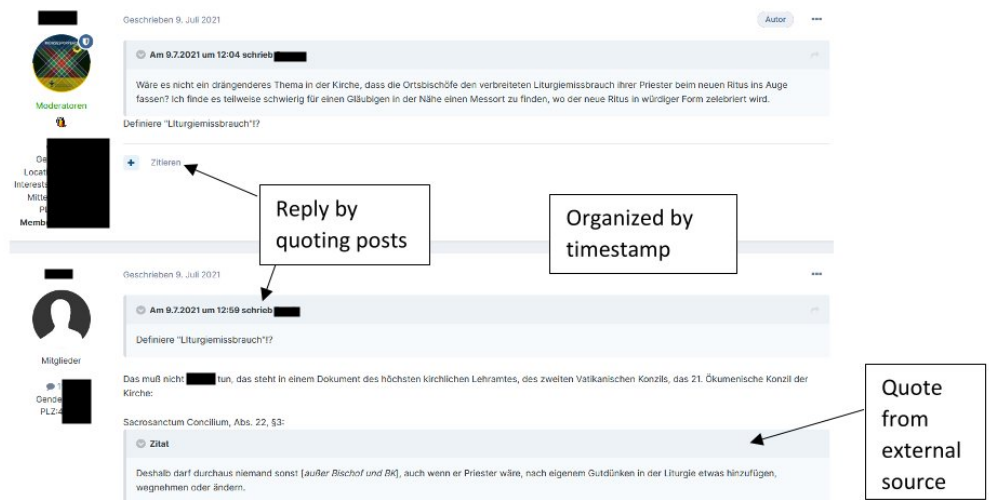
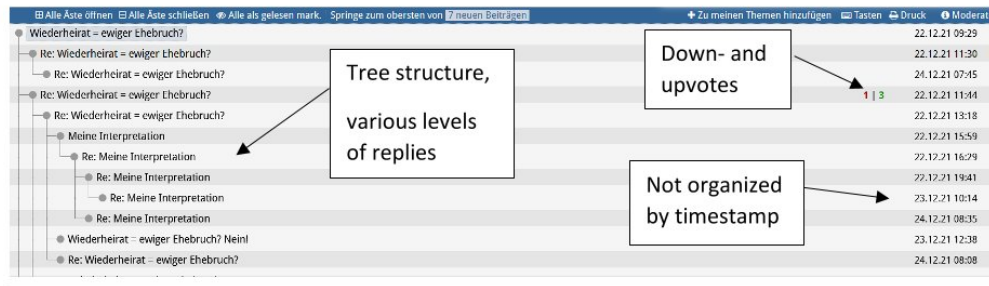


Fig. 2: Two German Christian online forums

```

<text>
  <body>
    <div type="thread">
      <post xml:id="p12902953" who="#user-2053702475" when="2021-12-22T09:29:00">
        <head>Wiederheirat = ewiger Ehebruch?</head>
        <fs type="post-reactions">
          <f name="upvotes">
            <numeric value="3"/>
          </f>
          <f name="downvotes">
            <numeric value="1"/>
          </f>
        </fs>
        <p>
          Ihr Lieben, ich bin seit Jahren glücklich verheiratet. Für meinen Mann und mich ist es nicht die erste Ehe. [...]
          [...] Sie meint, mein erster Mann sei mein wirklicher Mann und ich wäre verloren vor Gott, wenn ich mich nicht scheiden lasse [...]
        </p>
      </post>
    </div>
    <div type="replies">
      <div type="thread">
        <post xml:id="p12902973" replyTo="#p12902953" who="#user-851332273" when="2021-12-22T11:30:00">
          <head>Re: Wiederheirat = ewiger Ehebruch?</head>
          <p>
            <quote source="#p12902953"> Sie meint, mein erster Mann sei mein wirklicher Mann und ich wäre verloren vor Gott, wenn ich mich nicht scheiden lasse [...] </quote>
          </p>
          <p>Sie verlangt von dir die Scheidung, weil Scheidung Sünde ist... nun ja...</p>
        </post>
      </div>
      <div type="replies">
        <post xml:id="p12903412" replyTo="#p12902973" who="#user-753132775" when="2021-12-24T07:45:00">
          <head>Re: Wiederheirat = ewiger Ehebruch?</head>
          <p>
            <quote source="#p12902973"> Sie verlangt von dir die Scheidung, weil Scheidung Sünde ist... nun ja... </quote>
          </p>
          <p>
            genau so ist es [...]
          </p>
        </post>
      </div>
    </div>
  </div>
</body>

```

Fig. 3: Partial representation of a thread in TEI

Author Biographies

Sebastian Reimann is a research associate and PhD student in computational linguistics in the CRC 1475 "Metaphors of Religion" at the Center for Religious Studies at Ruhr University Bochum. His research interests include the computational analysis of texts in online communities and social media with a focus on the automatic detection of metaphors in religious online forums.

Lina Rodenhausen is a research associate and PhD student in religious studies in the CRC 1475 "Metaphors of Religion" at the Center for Religious Studies at Ruhr University Bochum where she works on religious communication in online forums. Her research focuses on digital religious communication and discourse analyses of contemporary religious communication as well as discourses on religion.

Frederik Elwert is the coordinator for digital humanities at the Center for Religious Studies at Ruhr University Bochum. His research interests include computational text analysis, network analysis, and modeling of cultural heritage data. A particular focus is the study of contemporary religious communication in online media. He is currently PI in the CRC 1475 "Metaphors of Religion."

Tatjana Scheffler is assistant professor for Digital Forensic Linguistics at Ruhr University Bochum, Germany. Her research focuses on analyzing conversations in digital media. She is a PI on several externally funded research projects on topics such as the variability of language in social media, computational analysis of metaphors in online forums, disinformation detection, and the semantics and pragmatics of emojis.

Short Paper: Publishing the grammateus research output with the TEI: how our scholarly texts become data

E. Nury

University of Geneva, Switzerland

Keywords: digital publications ; VRE ; open access ; scholarly communication ; web publication

Publishing the grammateus research output with the TEI: how our scholarly texts become data

Abstract:

The TEI is not exclusively used to encode primary sources: TEI-based scholarly publishing represents a non-negligible portion of TEI-encoded texts (Baillot and Giovacchini 2019). I present here how the encoding of secondary sources such as scholarly texts can benefit researchers, with the example of the grammateus project.

In the grammateus project, we are creating a Virtual Research Environment to present a new way of classifying Greek documentary papyri. This environment comprises a database of papyri, marked up with the standard EpiDoc subset of the TEI. It includes as well the textual research output from the project, such as introductory materials, detailed descriptions of papyri by type, and an explanation on the methodology of the classification. The textual research output was deliberately prepared as an online publication so as to fully take advantage of the interactivity with data offered by a web application, in contrast to a printed book. We are thus experimenting with a new model of scholarly writing and publishing. In this short paper I will describe how we have used the TEI not only for modeling papyrological data, but also for the encoding of scholarly texts produced in the context of the project, which would have traditionally been material for a monograph or academic articles. I will also demonstrate how this has enabled us later on to enrich our texts with markup for features that have emerged as relevant. We implemented a spiraling encoding process in which methodological documentation and analytical descriptions keep feeding back the editorial encoding of the scholarly texts. Documentation and analytical text therefore become data, within a research process based on a feedback method.

Bibliography:

Baillet, Anne, and Julie Giovacchini. 2021. 'TEI Models for the Publication of Social Sciences and Humanities Journals: Opportunities, Challenges, and First Steps Toward a Standardized Workflow'. *Journal of the Text Encoding Initiative*, no. Issue 14 (March).
<https://doi.org/10.4000/jtei.3419>.

Biography:

Elisa Nury is a postdoc at the University of Geneva and a scientific researcher at the Swiss Institute of Bioinformatics (DH+ group), where she is contributing to two Digital Humanities projects : grammateus and MARK16. She is interested in DH for classics, digital scholarly editions, and data visualization.

Short Paper: Handwritten Text Recognition for heterogeneous collections? The Use Case Gruß & Kuss

S. Büdenbender¹, M. Seltmann², J. Baum¹

¹: University of Applied Sciences Darmstadt (h_da), Germany; ²: University and State Library Darmstadt, Germany

Keywords: HTR, Transkribus, Citizen Science

Handwritten Text Recognition for heterogeneous collections? The Use Case Gruß & Kuss

Abstract

*Gruss & Kuss – Briefe digital. Bürger*innen erhalten Liebesbriefe*¹ – a research project funded by BMBF for 36 months – aims to digitize and explore love letters from ordinary persons with the

help of dedicated volunteers, also raising the question of how citizens can actively participate in the indexing and encoding of textual sources.

To present, transcriptions and basic annotations are made manually in Transkribus (lite), exported in page-xml and transformed into tei-xml, which serves as a basic format for archiving and web publication. Our corpus consists of more than 22,000 letters from 52 countries and 345 donators, divided into approximately 750 bundles (i.e., correspondences between usually two writers).² The oldest letter dates from 1715, the most recent from 2021, using a very broad concept of *letter* and including, for instance, notes left on pillows or WhatsApp messages.

The paper investigates the applicability of Handwritten Text Recognition (HTR) to this highly heterogeneous stock in a citizen science context. In an explorative approach, we will investigate at which scope of a bundle, respectively at which number of pages of the same handwriting, HTR becomes worthwhile.

For this purpose, the effort of a manual transcription is first compared to the effort of a model creation in Transkribus (in particular the creation of a training and validation set by double keying), including final corrections. In a second step, we will explore whether a modification of the procedure can be used to process even smaller bundles. Based on given metadata (time of origin, gender, script ...) a first clustering can be created, and existing models can be used as a basis for graphemically similar handwritings, allowing training sets to be kept much smaller while maintaining acceptable error rates. Another possibility is to start off with mixed training sets covering a class of related scripts.

Furthermore, we discuss how manual transcription by citizen scientists can be quantified in relation to the project's overall resources.

Bibliography

LBAKatalog (2022): Liebesbriefarchiv <http://katalog.liebesbriefarchiv.de>

Biographies

Jonathan Baum is a student research assistant at University of Applied Sciences Darmstadt (h_da), where he is studying Information Science M.Sc. with a focus on NLP, Semantic Web, and AI.

Stefan Büdenbender is a staff member at University of Applied Sciences Darmstadt (h_da) and works in the project *Gruß & Kuss* as well as in the NFDI Consortium Text+, Task Area Digital Editions. His areas of interest are retrodigitization of historical records, data modeling, and digital editions.

Melanie Seltmann currently works at the Centre for Digital Editions (CEiD) at the University and State Library Darmstadt in the project *Gruß & Kuss* as well as in the NFDI Consortium Text+, Task Area Digital Editions. She is interested in Digital Humanities, Citizen Science,

corpus linguistics, variational linguistics, standards & best practices, and annotation. In her PhD thesis she is focusing on annotations.

Notes:

1 *Love and kisses - <Encoding love letters in a citizen science approach>*

2 Cf. LBAKatalog 2022.

Session 4B – Long Papers – 09:30 - 11:00

Session 4B: Long Papers

Location: ARMB: 2.16

Chair: Elisa Beshero-Bondar, Penn State Behrend

Long Paper: From TEI Personography to IPIF data

R. W. J. Hadden¹, G. Vogeler^{2,1}

1: Austrian Academy of Sciences, Austria; 2: University of Graz, Austria

From TEI Personography to IPIF data

Richard Hadden¹ (richard.hadden@oeaw.ac.at); Georg Vogeler² (georg.vogeler@uni-graz.at)

¹Austrian Centre for Digital Humanities and Cultural Heritage, Vienna, Austria

²University of Graz, Austria

Keywords: IPIF, Prosopography, Personography, Linked Open Data

Brief Abstract

The International Prosopography Interchange Format (IPIF) is an open API and data model for prosopographical data interchange, access, querying and merging, using a regularised format. This paper discusses the challenges for converting TEI personographies into the IPIF format, and more general questions of using the TEI for so-called *factoid* prosopographies.

Biographies

Richard Hadden is a post-doctoral researcher in digital prosopography at the Austrian Centre for Digital Humanities and Cultural Heritage. His current work includes development of the IPIF standard, and he is the lead developer of the IPIF-Hub platform and other IPIF-related tools. He is also working on the DIGITARIUM project, applying machine learning techniques to information extraction from historical newspapers. He has a PhD in Digital Scholarly Editing from Maynooth University, Ireland.

Georg Vogeler is Chair for Digital Humanities at the *Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities* at the University of Graz. He is a trained historian who has worked at Ludwig-Maximilians-University, Munich; the Università del Salento, Lecce; the German Research Centre in Venice; the University of Vienna; and the Austrian Academy of Sciences. He has published on historical sciences, digital scholarly editing, semantic web technologies and digital prosopography

Full Abstract

Many TEI-based projects create considerable amount of person data, which (by and large) remain tightly bound to the edition or project in which it was created. As a solution to making such data available for external use, we present IPIF,¹ a REST API for prosopographical data, based on the *factoid prosopographical* model. The IPIF data model comprises four entities (a Factoid, linking a Person, a Source and a set of Statements), each queryable via a specific API endpoint. A proposal for IPIF-Hub (a centralised repository for IPIF data, modelled on the lines of *correspSearch*) is intended to further facilitate the use of the format.² We begin by presenting an overview of the API and the data model proposed by IPIF-Hub. However, mapping TEI personographies to IPIF presents several conceptual and technical difficulties, which we will address.

The TEI provides a rich set of facilities for describing person data,³ allowing more or less structured data to be represented, and closely integrated with encoded text. As demonstrated by Bradley & Jakacki's proposal,⁴ this enables TEI personographies to realise the *factoid* model (a 'factoid' is an assertion by a researcher that a source makes a given statement about a person: see Bradley 2005⁵). The simplest approach uses the <person> element as a basic construct, as it can handle complex prosopographical assertions: events, relationships, abstract traits and temporally determined states. Any of these elements may be freely duplicated, and explicitly marked with @resp and @source attributes – thus conforming to the factoid model (@source encodes the source of the assertion; @resp the researcher responsible). As such, the TEI should be considered as a useful approach to factoid prosopography, allowing database-like formalisms with a degree of flexibility afforded by its semi-structured nature.

However, using TEI for born-digital factoid prosopography (as opposed to encoding existing print-based prosopographies) is rare: the *Syriaca* project⁶ is the sole example of which we are aware. More typically, TEI personographies serve as adjuncts to digital scholarly editions, and do not follow a factoid approach: statements, characteristics, events, and persons are not marked with @source and @resp attributes; moreover, such statements are not necessarily formalised in a <person> element, but are simply references in the edition text. This scenario presents a number of conceptual difficulties for mapping such data to IPIF. Who is the "researcher" behind a given assertion? What is the source, if a property of the <person> is not based explicitly on a source? How can we determine the "text" of a statement when it is only implicitly referenced by, for instance, proximity to a <persName> element in the edition text? In this 'edition scenario', simple transformation to IPIF's factoid model is a challenge. While we present a number of common mappings for TEI elements into IPIF statements, it is clear that assumptions must be made: for instance, a factoid source and researcher responsibility can be inferred from the edition's metadata. However, this raises questions for the TEI on how to track the provenance of (in particular) annotations when these become more complex. We will address these difficulties, proposing a number of strategies, and discuss the feasibility of creating a generic transformation process. Finally, we will address the practicalities of ingesting converted data into the IPIF-Hub.

Notes:

1 International Prosopography Interchange Format (Vogeler et al. 2019),
https://d4h2020.sciencesconf.org/data/pages/Schlo_gl_Vogeler_Vasold_IPIF_2.pdf
See also: Vogeler, Georg, et al. 'Data Exchange in Practice: Towards a Prosopographical API'.
Proceedings of BD2019, edited by Antske Fokkens, 2019.

2 <https://github.com/IPIF/ipif-hub>

3 See Section 13 of the TEI guidelines:
<https://tei-c.org/release/doc/tei-p5-doc/en/html/ND.html> (accessed 2022/06/20).

4 Bradley, John & Diane Jakacki, <https://zenodo.org/record/6522540> (accessed 2022/06/20).

5 Bradley, John, 'Texts into Databases: The Evolving Field of New-Style Prosopography'.
Literary and Linguistic Computing, vol. 20, no. Suppl 1, Jan. 2005, pp. 3–24.
<https://doi.org/10.1093/lc/fqi022> (accessed 2022/06/20).

6 Schwartz, Daniel L., et al. 'Modeling a Born-Digital Factoid Prosopography Using the TEI and Linked Data'. *Journal of the Text Encoding Initiative*, 2020, p. 37.
<https://journals.openedition.org/jtei/3979> (accessed 2022/06/20).

Long Paper: TEI as Data: Escaping the Visualization Trap

R. Rosselli Del Turco¹, E. Magnanti², G. Cerretini³

1: Università di Torino, Italy; 2: University of Vienna, Austria; 3: Università di Pisa, Italy

Keywords: data modeling, information retrieval, data processing, digital philology, digital editions

TEI as Data: Escaping the Visualization Trap

Giacomo Cerretini, Università di Pisa, Italia - cerre.giacomo93@gmail.com

Elisabetta Magnanti, University of Vienna, Austria - elisabetta.magnanti@univie.ac.at

Roberto Rosselli Del Turco, Università di Torino, Italia - roberto.rossellidelturco@unito.it

Abstract

During the last few years, the TEI Guidelines and schemas have continued growing in terms of capability and expressive power. A well-encoded TEI document constitutes a small treasure trove of textual data that could be queried to quickly derive information of different types. However, access to such data is mainly intended for visualization purposes in many edition browsing tools, e.g. EVT (<http://evt.labcd.unipi.it/>). Such an approach seems to be hardly compatible with the strategy of setting up databases to query this data, thus leading to a splitting of environments: DSEs to browse edition texts versus databases to perform powerful and sophisticated queries. It would be interesting to expand the capabilities of EVT, and possibly other tools, adding functionalities which would allow them to process TEI documents to answer complex user queries. This requires both an investigation to define the text model in terms of TEI elements and a subsequent implementation of the desired functionality, to be tested on a suitable TEI project that can adequately represent the text model.

The *Anglo-Saxon Chronicle* stands out as an ideal environment to test such a method. The wealth of information that it records about early medieval England makes it the optimal footing upon which to enhance computational methods for textual criticism, knowledge extraction and data modeling for primary sources. The application of such a method could here prove essential to assist the retrieval of knowledge otherwise difficult to extract from a text that survives in multiple versions. Bridging together, cross-searching and querying information dispersed in all the witnesses of the tradition would allow us to broaden our understanding of the *Chronicle* in unprecedented ways. Interconnecting the management of a wide spectrum of named entities and realia—which is one of the greatest assets of TEI—with the representation of historical events would make it possible to gain new knowledge about the past. Most importantly, it would lay the groundwork for a Digital Scholarly Edition of the *Anglo-Saxon Chronicle*, a project never undertaken so far.

Therefore, we decided to implement a new functionality capable of extracting and processing a greater amount of information by cross-referencing various types of TEI/XML-encoded data. We developed a TypeScript library to outline and expose a series of APIs allowing the user to perform complex queries on the TEI document. Besides the cross referencing of people, places and events as hinted above—on the basis of standard TEI elements such as `<listPerson>/<person>`, `<listPlace>/<place>`, `<listEvent>/<event>` etc.—we plan to support ontology-based queries, defining the relationships between different entities by means of RDF-like triples. In a similar way, it will be possible to query textual variants recorded in the critical apparatus by typology and witness distribution. This library will be integrated in EVT to interface directly with its existing data structures, but it is not limited to it. We are currently working on designing a dedicated GUI within EVT to make the query system intuitive and user-friendly.

The research and development work leading to this paper originates from an international workshop held in Pisa in June 2020 (*Medieval Archival Sources into the Digital. The Challenge of Processing and Visualising Semi-structured Data* <http://www.labcd.unipi.it/fonti-archivistiche-medievali-nel-digitale/>). This initial discussion was later expanded with the latest research about supporting ontologies in TEI (see in particular <https://github.com/TEIC/TEI/issues/1860>) and in the next EVT release (“There and back again: what to expect in the next EVT version”: <http://amsacta.unibo.it/6848/>). The paper will also deal with development strategies aimed at simplifying the implementation of new features in the forthcoming versions of EVT.

References

- Cacioli, Giulia, Giacomo Cerretini, Chiara Di Pietro, Sara Maenza, Roberto Rosselli Del Turco and Simone Zenzaro. 2022. “There and back again: what to expect in the next EVT version”. In Fabio Ciraci, Giulia Miglietta, Carola Gatto (eds.), *AIUCD 2022 - Digital cultures. Intersections: philosophy, arts, media. Proceedings of the 11th national conference, Lecce, 1-3 June 2022*: 212-217. <http://amsacta.unibo.it/6848/>.

- Dumville, David and Simon Keynes, gen. eds. *The Anglo-Saxon Chronicle: A Collaborative Edition*. Cambridge: D.S. Brewer, 1983–.
- Keynes, Simon. “Manuscripts of the Anglo-Saxon Chronicle.” In *The Cambridge History of the Book in Britain*, 1:537–52. Cambridge: Cambridge University Press, 2011. <https://doi.org/10.1017/CHOL9780521583459.026>.
- Rosselli Del Turco, Roberto. 2019. “Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology).” *Textual Cultures* 12 (2): 91–111. <https://doi.org/10.14434/textual.v12i2.27690>.
- ———. 2021. “Elaborazione di dati semi-strutturati: ipotesi implementative e casi d’uso tratti da testi in inglese antico.” *Umanistica Digitale*, no. 10 (September): 387–407. <https://doi.org/10.6092/issn.2532-8816/12598>.
- Rosselli Del Turco, Roberto, Enrica Salvatori, Andrea Nanetti, Marco Giacchetto, Vera Isabell Schwarz-Ricci, and Antonella Ambrosio. 2021. “Introduzione: ‘Fonti archivistiche medievali nel digitale. La sfida di trattare e visualizzare dati semi-strutturati.’” *Umanistica Digitale*, no. 10 (September): 289–98. <https://doi.org/10.6092/issn.2532-8816/12582>.
- Thaller, Manfred. 2020. “Can historical information be represented outside of a graph / hypergraph / network?” In *Graph Technologies in the Humanities: 2021 Virtual Symposium*, <https://graphentechnologien.hypotheses.org/files/2021/02/Thaller-Mainz2021-2.pdf>.
- Vogeler, Georg. 2019. “The ‘Assertive Edition’: On the Consequences of Digital Methods in Scholarly Editing for Historians.” *International Journal of Digital Humanities* 1 (2): 309–22. <https://doi.org/10.1007/s42803-019-00025-5>.
- Whitelock, Dorothy (transl.) with David C. Douglas and Susie I. Tucker. 1961. *The Anglo-Saxon Chronicle. A Revised Translation*. London: Eyre and Spottiswoode.

Biographies

Giacomo Cerretini is a student in the Master’s degree program in Digital Humanities at the University of Pisa. Over time, Giacomo has developed skills and experience as a front-end and back-end developer through collaboration with the EVT project, with the Net7 DH developers and various other university-based projects. In addition to his main job functions, Giacomo also devotes time to the study of application fields such as network and data analysis.

Elisabetta Magnanti is a University Assistant and Doctoral Candidate at the Department of History, at the University of Vienna. Her research specialises in early medieval England and in the application of computational methods to philological and historical analysis of Old English source material. She also contributes to the ongoing *Digital Vercelli Book* project.

Roberto Rosselli Del Turco is an Associate Professor at the Università di Torino, Dipartimento di Studi Umanistici, where he teaches Germanic Philology and Digital Philology. He has published widely in the Digital Humanities and Anglo-Saxon fields of study. He is the director of the *Digital Vercelli Book* project and co-director of the Visionary Cross project. He also is the director of Edition Visualization Technology, a DSE browsing software (<http://evt.labcd.unipi.it/>).

Long Paper: LINCS' Linked Workflow: Creating CIDOC-CRM from TEI

C. Crompton, H. Zafar, A. Defours

University of Ottawa, Canada

Keywords: linked data, conversion, reconciliation, software development

LINCS' Linked Workflow: Creating CIDOC-CRM from TEI

TEI data is so often carefully curated without any of the noise and error common to algorithmically created data, that it is a perfect candidate for linked data creation; however, while most small TEI projects boast clean beautifully crafted data, linked data creation is often out of reach both technically and financially for these project teams. This paper reports (following where others have tread¹) on the Networked Cultural Scholarship project (LINCS) workflow, mappings, and tools for creating linked data from TEI resources.

The process of creating linked data is far from straightforward since TEI is by nature hierarchical, taking its meaning from the deep nesting of elements. Any one element in TEI may be drawing its meaning from its relationship to a grandparent well up the tree (for example a `persName` appearing inside a `listPerson` inside the `teiHeader` is more likely to be a canonical reference to a person than a `persName` whose parent is a paragraph). Furthermore, the meaning of TEI elements are not always well-represented in existing ontologies and the time and money required to represent TEI-based information about people, places, time, and cultural production as linked data is out of reach of many small projects.

This paper introduces the LINCS workflow for creating linked data from TEI. We will introduce the named entity recognition and reconciliation service, NSSI (pronounced *nessy*), and its integration into a TEI-friendly vetting interface, Leaf Writer. Following NSSI reconciliation, Leaf Writer users can download their TEI with the entity `uris` in `idno` elements for their own use. If they wish to contribute to LINCS, they may proceed to enter the TEI document they have exported from Leaf Writer into XTriples, a customized version of Mainz's Digitale Akademie's XTriples tool of the same name, which converts TEI to CIDOC-CRM for either private use, or for integration into the LINCS repository. We have adopted the XTriples tool because it meets the needs of a very common type of TEI user: the director or team member of a project who is not going to be able to learn the intricacies of CIDOC-CRM, or indeed perhaps not even of linked data principles, but would still like to contribute their data to LINCS. That said, we are keen to get the feedback of the expert users of the TEI community on our workflow, CIDOC-CRM mapping, and tools.

Biographies

Constance Crompton is a Canada Research Chair in Digital Humanities at the University of Ottawa with research interests in linked data, data modelling, queer history, and Victorian popular culture.

Huma Zafar is a developer at the University of Ottawa library, and a master's student in the School of Information Studies, and lead on NSSI and LINCS' XTriples.

Alice Defours is a PhD student in Communication at the University of Ottawa and a research assistant for the LINCS project, with a particular focus on XSLT-based conversion.

Notes:

1 Bodard, Gabriel, Hugh Cayless, Pietro Liuzzo, Chiara Cenati, Alison Cooley, Tom Elliott, Silvia Evangelisti, Achille Felicetti, et al. "Modeling Epigraphy with an Ontology." *Zenodo*, March 26, 2021. <https://doi.org/10.5281/zenodo.4639508>.

Ciotti, Fabio. "A Formal Ontology for the Text Encoding Initiative." *Umanistica Digitale*, vol. 2, no. 3, 2018.

Eide, Ø., and C. Ore. "From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration Between Text Collections and Other Sources of Cultural Historical Documentation." *Digital Humanities*, 2007.

Liuzzo, Pietro, et al. "Networking EAGLE with CIDOC and TEI." *ICOM: Access and Understanding – Networking in the Digital Era*, 2014/

Ore, Christian-Emil, and Øyvind Eide. "TEI and Cultural Heritage Ontologies: Exchange of Information?" *Literary and Linguistic Computing*, vol. 24, no. 2, 2009, pp. 161-72., <https://doi.org/10.1093/llc/fqp010>.

Session 5A – Long Papers – 11:30 - 13:00

Session 5A: Long Papers

Location: ARMB: 2.98

Chair: Dario Kampkaspar, Universitäts- und Landesbibliothek Darmstadt

Long Paper: Evolving Hands: HTR and TEI Workflows for cultural institutions

J. Cummings¹, D. Jakacki², I. Johnson¹, C. Pirmann², A. Healey¹, V. Flex¹, E. Jeffrey¹

¹: Newcastle University, United Kingdom; ²: Bucknell University, USA

Keywords: TEI XML, Handwritten Text Recognition, HTR, Libraries

TEI2022 – Evolving Hands: HTR and TEI Workflows for cultural institutions

This Long Paper will look at the work of the Evolving Hands project which is undertaking three case studies ranging across document forms to demonstrate how TEI-based HTR workflows can be iteratively incorporated into curation. These range from: 19th-20th century handwritten letters and diaries from the UNESCO Gertrude Bell Archive, 18th century German, 20th century French correspondence, and a range of printed materials from the 19th century onward in English and French. A joint case study converts legacy printed material of the Records of Early English Drama (REED) project. By covering a wide variety of periods and document forms the project has a real opportunity here to foster responsible and responsive support for cultural institutions.

Newcastle University Case Study -- The Gertude Bell Archive

This case study uses Newcastle Special Collection's UNESCO Gertrude Bell Archive (<http://gertrudebell.ncl.ac.uk/>), which document the activities of the explorer, archaeologist, and political agent who was instrumental in establishing the Kingdom of Iraq in 1921. Bell is the subject of plays, documentaries, a feature film, and recently was nominated as a BBC 20th Century Icon. A separate centenary project is digitizing and cataloguing her archive of diaries, letters, and photographs. Piggybacking on that, we will select the richest materials to train HTR base models of Bell's hand and use up-converted transcriptions for the production of training materials.

Bucknell University Case Study -- Scholarly Production at Scale

The Bucknell case study centres on processes used across multiple discrete projects by staff with a range of digital experience. These projects represent different models for testing the HTR to TEI conversion process. Their sources are drawn from Bucknell's Special Collections and research of faculty working with archives in the US, UK, Europe, and Asia. They include scribal hands, life papers, correspondence, and semi-legible typed government files from 1700-1990 and are in English, French, German, and Maithili. This case study will directly benefit multiple projects at the university, and is optimised for sharing with smaller cultural institutions around the world.

Joint Newcastle/Bucknell Case Study -- Transforming REED Print Collections

Cummings (AHRC PI) and Jakacki (NEH PI) will collaborate on a case study converting collections produced by the Records of Early English Drama (<http://reed.utoronto.ca>) project that has published since 1979 edited documentary records of pre-1642 performance in premodern England, Scotland and Wales. However, the semantic information provided in the print collections, through the use of special symbols and formatting, is lost in OCR. Previous tests using HTR by Jakacki and Cummings have demonstrated that these distinctions can be transformed with HTR to TEI. The project will document shared workflows for consistent upconversion into viable materials ready to enter the REED project's digital publication workflow. This has the potential to be of use for all the other REED legacy print volumes (well over 20,000 pages of rich scholarly material).

Biographies:

- James Cummings:
 - James Cummings is the Senior Lecturer in Late Medieval Literature and Digital Humanities for the School of English Literature, Language, and Linguistics at Newcastle University. He is the Newcastle PI for the Evolving Hands Project, the TEI Board of Directors, and for some reason volunteered to be lead local organiser for the TEI2022 conference.
- Diane Jakacki:
 - Diane Jakacki is Digital Scholarship Coordinator and Associated Faculty in Comparative & Digital Humanities at Bucknell University. She is PI of the Mellon-funded LAB Cooperative project, co-PI (with James Cummings) of the Evolving Hands project, and member of the LEAF executive team. She is chair of the TEI Board of Directors and Chair-Elect of the Alliance for Digital Humanities Organizations.
- Ian Johnson:
 - Ian Johnson is Head of Special Collections & Archives at Newcastle University Library with overall responsibility for the curation of our unique and distinctive collections for collaborative research, teaching and engagement. This includes interdisciplinary digital scholarship and co-curation of our UNESCO Gertrude Bell Archive. He is co-I for the Evolving Hands project.
- Carrie Pirmann:
 - Carrie Pirmann is the social sciences librarian at Bucknell University who is optimising existing Transkribus models for the Bucknell Case Study and working with the digital libraries community.
- Alexandra Healey:
 - Alexandra Healey is a Project Archivist in Newcastle University Special Collections & Archives. She is coordinating the use of HTR and TEI within the Newcastle team as part of the Evolving Hands project. She also designs and delivers teaching around the use of archives in digital scholarship and is Chair of the Archives for Learning and Education Section of the Archives and Records Association.
- Valentina Flex:
 - Valentina is the Stillman Project Archivist working on Gertrude Bell Archive: Bell and the Kingdom of Iraq at 100.

- Evie Jeffrey:
 - Evie Jeffrey is the postgraduate assistant for the project and started working on the project as a Robinson Bequest Bursary-holder.

Long Paper: Between automatic and manual encoding: towards a generic TEI model for historical prints and manuscripts

A. Pinche¹, K. Christensen², S. Gabay³

1: Ecole nationale des chartes | PSL (France); 2: INRIA (France); 3: Université de Genève (Switzerland)

Keywords: TEI, text extraction, linguistic annotation, digital edition, mass digitisation

Between automatic and manual encoding: towards a generic TEI model for historical prints and manuscripts

Ariane Pinche (École nationale des chartes | PSL, France)

Kelly Christensen (INRIA, France)

Simon Gabay (Université de Genève, Switzerland)

Abstract

Cultural heritage institutions today aim to digitise their collections of prints and manuscripts (Bermès 2020) and are generating more and more digital images (Gray 2009). To enrich these images, many institutions work with standardised formats such as IIIF, preserving as much of the source's information as possible. To take full advantage of textual documents, an image alone is not enough. Thanks to automatic text recognition technology, it is now possible to extract images' content on a large scale. The TEI seems to provide the perfect format to capture both an image's formal and textual data (Janès et al. 2021). However, this poses a problem. To ensure compatibility with a range of use cases, TEI XML files must guarantee IIIF or RDF exports and therefore must be based on strict data structures that can be automated. But a rigid structure contradicts the basic principles of philology, which require maximum flexibility to cope with various situations.

The solution proposed by the *Gallic(orpor)a* project¹ attempted to deal with such a contradiction, focusing on French historical documents produced between the 15th and the 18th c. It aims to enrich the digital facsimiles distributed by the French National Library (BnF)² in two different ways:

- text extraction, including the segmentation of the image (layout analysis) with *SegmOnto* (Gabay, Camps, et al. 2021) and the recognition of the text (Handwritten Text Recognition) augmenting already existing models such as Pinche and Clérice (2021);
- linguistic annotation, including lemmatisation, POS tagging (Gabay, Clérice, et al. 2020), named entity recognition and linguistic normalisation (Bawden et al. 2022).

Our TEI document modelling has two strictly coercive automatically generated data blocks:

- the <sourceDoc> with information from the digital facsimile, which computer vision, HTR and segmentation tools produce thanks to machine learning (Scheithauer et al. 2021);

- the <standOff> (Bartz et al. 2021a) with linguistic information produced by natural language processing tools (Gabay, Suarez, et al. 2022) to make it easier to search the corpus (Bartz et al. 2021b).

Two other elements are added that can be customised according to researchers' specific needs:

- a pre-filled <teiHeader> with basic bibliographic metadata automatically retrieved from (i) the digital facsimile's IIF Image API and (ii) the BnF's Search/Retrieve via URL (SRU) API³. The <teiHeader> can be enriched with additional data, as long as it respects a strict minimum encoding;
- a pre-editorialised <body> (fig. 1). It is the only element totally free regarding encoding choices.

```
<body>
  <div>
    <pb corresp="#page5"/>
    <note corresp="#page5_zone2" type="MarginTextZone">
      <lb corresp="#page5_zone2_line1"/>79/4120
    </note>
    <pb corresp="#page6"/>
    <ab corresp="#page6_zone1" type="MainZone">
      <hi rend="HeadingLine">
        <lb corresp="#page6_zone1_line1"/>BRADAMANTE,
        <lb corresp="#page6_zone1_line2"/>TRAGECOMDEDIE.
      </hi>
    </ab>
    <pb corresp="#page9"/>
    <fw corresp="#page9_zone1" type="RunningTitleZone">
      <lb corresp="#page9_zone1_line1"/>AV ROY.
    </fw>
    <ab corresp="#page9_zone2" type="MainZone">
      <lb corresp="#page9_zone2_line1"/>uiuront nostre siecle, les admira-
      <lb corresp="#page9_zone2_line2"/>bles effets de vos heroiques ver-
      <gap reason="sampling"/>
    </ab>
  </div>
</body>
```

Figure 1: Example of a pre-editorialised <body>: Robert Garnier, *Tragédies*, Paris: Robert Estienne, 1582 [ark:/12148/bpt6k990549b].

By restricting certain elements and allowing others to be customisable, our TEI model can efficiently pivot toward other export formats, including RDF and IIF. Furthermore, the <sourceDoc> element's strict and thorough encoding of all of the document's graphical information allows the TEI document to be converted into PAGE XML and ALTO XML files, which can then be used to train OCR, HTR, and segmentation models. Thus, not only does our TEI model's strict encoding avoid limiting philological choices, thanks to the <body>, it also allows us to pre-editorialise the <body> via the content of the <sourceDoc> and, in a near future, the <standOff>.

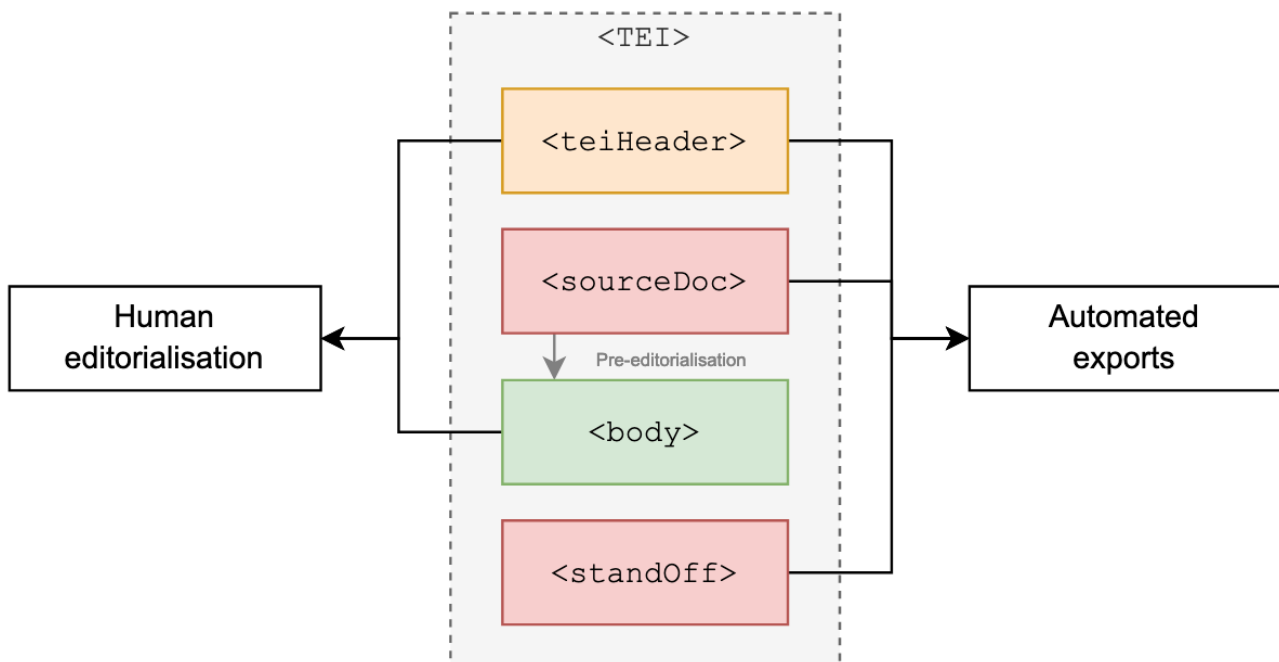


Figure 2: Gallic(orpora) TEI model. In red strictly coercive automatically generated data. In green fully customisable data. In orange partially customisable data.

Data and Script

Data and scripts are available online: <https://github.com/Gallicorpora>.

References

- Bartz, A. et al. (Oct. 2021). “Expanding the content model of annotationBlock”. In: *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Virtual, United States. url: <https://hal.archives-ouvertes.fr/hal-03380805>.
- Bawden, R. et al. (June 2022). “Automatic Normalisation of Early Modern French”. In: *LREC 2022-13th Language Resources and Evaluation Conference*. European Language Resources Association. Marseille, France. doi: 10.5281/zenodo.5865428. url: <https://hal.inria.fr/hal-03540226>.
- Bermès, E. (Jan. 25, 2020). “Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)”. These de doctorat. Paris, École nationale des chartes. url: <http://www.theses.fr/2020ENCP0001>
- Gabay, S., J. -B. Camps, et al. (Sept. 2021). “SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)”. In: *Proceedings of the 1st International Workshop on Computational Paleography, IWCP@ICDAR 2021*. Lausanne (Switzerland): Springer.
- Gabay, S., T. Clérice, et al. (Oct. 2020). “Standardizing linguistic data: method and tools for annotating (pre-orthographic) French”. In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia. doi: 10.1145/3423603.3423996. url: <https://hal.archives-ouvertes.fr/hal-03018381>.
- Gabay, S., P. O. Suarez, et al. (June 2022). “From FreEM to D'AlembERT”. <https://hal.inria.fr/hal-03596653>.
- Gray, J. (2009). “Jim Gray on eScience: A transformed scientific method”. In: *The fourth paradigm: Data-intensive scientific discovery*. Ed. by T. Hey, S. Tansley, and K. Tolle.

Washington, pp. xvii – xxxi.

Janès, J. et al. (Dec. 2021). “Towards automatic TEI encoding via layout analysis”. In: *Fantastic future 21, 3rd International Conference on Artificial Intelligence for Libraries, Archives and Museums*. Paris, France: AI for Libraries, Archives, and Museums (ai4lam). url: <https://hal.archives-ouvertes.fr/hal-03527287>.

Pinche, A. and T. Clérice (Aug. 2021). HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI). Version 1.0.1. doi: 10.5281/zenodo.5235186. url: <https://doi.org/10.5281/zenodo.5235186>.

Scheithauer, H. et al. (Oct. 2021). “From page to content – which TEI representation for HTR output?” In: *Next Gen TEI, 2021 – TEI Conference and Members’ Meeting*.

Weaton (virtual), United States. url: <https://hal.archives-ouvertes.fr/hal-03380807>.

Biographies

1. Ariane Pinche is a postdoctoral fellow at the École nationale des chartes | PSL and currently works on Medieval manuscripts and HTR in CREMMALab and *Gallic(orpor)a* projects.
2. Kelly Christensen is an intern at Inria and a member of the *Gallic(orpor)a* project. She holds a PhD in musicology and is specialised in digital humanities.
3. Simon Gabay is a senior lecturer and researcher at the university of Geneva. He is specialised in Early Modern French literature and participates in the *Gallic(orpor)a* project.

1<https://gallicorpora.github.io>.

2<https://gallica.bnf.fr>.

3<https://catalogue.bnf.fr/api>

Long Paper: Dehmel Digital: Pipelines, text as data, and editorial interventions at the distance

D. Maus¹, J. Nantke², S. Bläfs², M. Flüh²

1: State and University Library Hamburg, Germany; 2: University of Hamburg

Keywords: NER, HTR, Correspondence, Digital Scholarly Edition

Dehmel Digital: Pipelines, text as data, and editorial interventions at the distance

David Maus (State and University Library Hamburg), Julia Nantke (University of Hamburg), Sandra Bläfs (University of Hamburg), Marie Flüh (University of Hamburg)

Ida and Richard Dehmel were a famous, internationally well-connected artist couple around 1900. The correspondence of the Dehmels, which has been comprehensively preserved in approx. 35,000 documents, has so far remained largely unexplored in the Dehmel Archive of the State and University Library Hamburg. The main reason for this is the quantity of the material that makes it difficult to explore the material using traditional methods of scholarly editing. However, the corpus is relevant for future research precisely because of its size and variety. It not only contains many letters from important personalities from the arts and culture of the turn of the century, but also documents personal relationships, main topics as

well as forms and ways of communication in the cultural life of Germany and Europe before the First World War on a large scale.

The project *Dehmel digital*¹ seeks to close this gap by creating a digital scholarly edition of the Dehmels' correspondence that addresses the quantitative aspects with a combination of state-of-the-art machine learning approaches, namely handwritten text recognition (HTR) and named entity recognition (NER). At the heart of the project is a scalable pipeline that integrates automated and semi-automated text/data processing tasks. In our paper we will introduce and discuss the main steps: 1. Importing the result of HTR from Transkribus and OCR4all, 2. Applying a trained NER model; 3. Disambiguating entities and referencing authority records with OpenRefine; 4. Publishing data and metadata to a Linked Open Data web service. Our main focus will be on the pipeline itself, the “glue” that ties together well-established tools (Transkribus, OCR4All, Stanford Core NLP, OpenRefine), our use of TEI to encode relevant information and the special challenges we observe when using text as data, i.e. combining automated and semi-automated processes with the desire of editorial interventions.

¹ <https://dehmel-digital.de>

Session 5B – Panel: Manuscript catalogues as data for research – 11:30 - 13:00

Session 5B: Panel - Manuscript catalogues as data for research

Location: ARMB: 2.16

Chair: Katarzyna Anna Kapitan, University of Oxford

Panel: Manuscript catalogues as data for research

H. E. Jones¹, Y. Faghihi¹, M. Holford², T. Schaßan³, T. Burrows², K. A. Kapitan², N. K. Yavuz⁴

1: Cambridge University, United Kingdom; 2: University of Oxford; 3: Herzog August

Bibliothek; 4: University of Leeds

Keywords: Manuscripts, Provenance, Research, Clustering, Linked Data

Manuscript catalogues present problems and opportunities for researchers, not least the status of manuscript descriptions as both information about texts and texts in themselves. In this panel, we will present three recent projects which have used manuscript catalogues as data for research, and which raise general questions in text encoding, in manuscript studies and in data-driven digital humanities. This will be followed by a panel discussion to further investigate issues and questions raised by the papers.

Investigating the Origins of Islamicate Manuscripts Using Computational Methods (Yasmin Faghihi and Huw Jones):

This project evaluated computational methods for the generation of new information about the origins of manuscripts from existing catalogue data. The dataset was the Fihrist Union Catalogue of Manuscripts from the Islamicate World, which contains c.13,000 TEI manuscript descriptions from the collections of 21 UK institutions. We derived a set of codicological features from the TEI data, clustered together manuscripts sharing features, and used dated/placed manuscripts to generate hypotheses about the provenance of other manuscripts in the clusters. We aimed to establish a set of base criteria for the dating/placing of manuscripts, to investigate methods of enriching existing datasets with inferred data to form the basis of further research, and to engage critically with the research cycle in relation to computational methods in the humanities.

Re-thinking the <provenance> element in TEI Manuscript Description to support graph database transformations (Toby Burrows and Matthew Holford):

This paper reports on the transformation of the Bodleian Library's online medieval manuscripts catalogue, based on the "Manuscript Description" section of the TEI Guidelines, into RDF graphs using the CIDOC-CRM and FRBR_{oo} ontologies. This work was carried out in the context of two Linked Open Data projects: Oxford Linked Open Data (OxLOD) and Mapping Manuscript Migrations (MMM).

One area of particular focus was the provenance data relating to these manuscripts, which proved challenging to transform effectively from TEI to RDF. An important output from the MMM project was a set of recommendations for re-thinking the structure and encoding of the

TEI <provenance> element to enable more effective reuse of the data in graph database environments. These recommendations draw on concepts previously outlined by Ore and Eide (2009), but also take into account the parallel work being done in the art museum and gallery community on documenting and reusing provenance information (Bergen-Fulton, Newbury, and Snyder 2015; Knoblock et al. 2017).

The use of TEI in the Handschriftenportal (Torsten Schaßan)

The national manuscript portal for Germany in the making, the Handschriftenportal, is built on TEI encoded data. These include representations for manuscripts, descriptions that have been imported, authority data, and OCR-generated catalogues. In the future, it will be possible to enter descriptions directly into the backend database.

The structure of the descriptive data shall be adopted according to the latest developments in manuscript studies, e.g. the risen importance of material aspects, or the alignment of the description of texts and illuminations.

Especially the latter, the data to be entered in the future, poses several issues to the TEI encoding as currently defined in the Guidelines. This comprises the overall structure of the main components of a description such as <msContents>, <physDesc>, and <msPart>s, as well as needs on a more detailed level such as it is currently not possible to enter data on musical contents on the <msItem> level. In the paper our approaches towards and needs for change of TEI encoding will be presented.

Biographies:

Toby Burrows: Dr Toby Burrows is a Digital Humanities researcher at the University of Oxford and the University of Western Australia. His research focuses on the history of cultural heritage collections, and especially medieval and Renaissance manuscripts. His recent projects include: Collecting the West, Mapping Manuscript Migrations, and HuNI: the Australian Humanities Networked Infrastructure. He is also a contributor to the AHRC DigiSpec project.

Yasmin Faghihi: Yasmin Faghihi is Head of the Near and Middle Eastern Department at Cambridge University Library. She is the editor of FIHRIST, the online union catalogue for manuscripts from the Islamic world. She has been leading on standardised practices in text encoding for manuscript description and teaching to foster awareness about compatible approaches to data creation and use.

Matthew Holford: Matthew Holford is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, Oxford. He has a long-standing research interest in the use of TEI for the description and cataloguing of Western medieval manuscripts.

Huw Jones: Huw Jones is Head of the Digital Library at Cambridge University Library, and Director of CDH Labs at Cambridge Digital Humanities. His work spans many aspects of

collections-driven digital humanities, from creating and making collections available to their use in a research and teaching context.

Torsten Schaßan: Torsten Schaßan is member of the Manuscripts and Special Collections department of the Herzog August Bibliothek Wolfenbüttel. He was involved in many manuscript digitisation and cataloguing projects. In the Handschriftenportal project he is responsible for the definition of schemata and all transformations of data for import into the portal.

Chair: Dr Katarzyna Anna Kapitan is manuscript scholar and digital humanist specialising in Old Norse literature and culture. Currently she is Junior Research Fellow at Linacre College, University of Oxford, where she works on a digital book-historical project, “Virtual Library of Torfæus”, funded by the Carlsberg Foundation. She published on applications of DH to manuscript studies and taught DH courses in fundamentals of TEI-XML, digital scholarly editing and cataloguing as well as computer assisted textual criticism.

Respondent: Dr N. Kivılcım Yavuz works at the intersection of medieval studies and digital humanities, with an expertise in medieval historiography and European manuscript culture. She is especially interested in digitisation of manuscripts as cultural heritage items and creation, collection and interpretation of data and metadata in the context of digital repositories. She is a member of the Executive Board of Digital Medievalist and serves on the MDR (Database of Medieval Digital Resources) Committee of the Medieval Academy of America. <https://nkyavuz.com/biography/>

Session 6A – An Interview With ... Lou Burnard – 14:30 - 16:00

Session 6A: An Interview With ... Lou Burnard

Time: Thursday, 15/Sept/2022: 2:30pm - 4:00pm

Session Chair: Diane Jakacki, Bucknell University

Location: ARMB: 2.98

TEI2022 introduces a new conference session format: “An Interview With...” where a prominent member of the TEI is interviewed. After an introduction, the interviewee will make a short statement piece, followed by interview questions, which are then opened up to the audience.

This year’s inaugural interviewee is Lou Burnard, previously one of the editors of the TEI-C Guidelines.

TEI Consortium Annual General Meeting – 16:30 - 18:00

TEI Annual General Meeting - All Welcome

Time: Thursday, 15/Sept/2022: 4:30pm - 6:00pm

Session Chair: Diane Jakacki, Bucknell University

Location: ARMB: 2.98

The TEI Constortium Annual General Meeting will be held from 4:30-6:00pm on Thursday 15 September 2022 at the TEI2022 Conference. All attendees of the conference are welcome to come and hear reports on the activities of the TEI Consortium over the last year.

Conference Sessions – Friday 16 September 2022

Session 7A – Short Papers – 09:30 - 11:00

Session 7A: Short Papers

Location: ARMB: 2.98

Chair: Patricia O Connor, University of Oxford

Short Paper: Encoding Complex Structures: The Case of a Gospel Spanish Chapbook

E. Leblanc, P. Jacsont

University of Geneva, France

Keywords: Spanish literature, Digital library, TEI-Publisher, facsimile, sourceDoc

Encoding Complex Structures: The Case of a Gospel Spanish Chapbook

The project *Untangling the cordel* seeks to study and revalue a corpus of Spanish chapbooks dating from the 19th century by creating a digital library (Leblanc and Carta 2021). This corpus of chapbooks, also called pliegos de cordel, is highly heterogeneous in its content and editorial formats, giving rise to multiple reflections on its encoding.

In this short paper, we would like to share our feedback and thoughts on the XML-TEI encoding of a Gospel pliego for its integration into TEI-Publisher.

This pliego is an in-4° containing 16 small columns with extracts from the Four Gospels (John's prologue, Annunciation, Nativity, Mark's finale and the passion according to John; i.e. the same extracts as those in the book of hours (Join-Lambert 2016)) duplicated on both sides. This printout had to be cut in half and then folded to obtain two identical sets of excerpts from the Four Gospels. Whoever acquires it appropriates the object for private devotions or protection: it is therefore not an object kept for reading (the text is written in Latin with small letters) but for apotropaic or curative use (Botrel 2021).

To put forward the interest of this pliego as a devotional object and not strictly as a textual object required much reflection concerning its encoding and its publication on our digital library. Indeed, depending on our choice of encoding, the information conveyed differs: should we favour a diplomatic and formal edition or an encoding that follows the reading?

To determine which encoding would be the most suitable, we decided to test two encoding solutions, one with <facsimile> and another with <sourceDoc>. The visualisation of the two encodings possibilities on TEI-Publisher will allow us to develop the advantages and disadvantages of each method.

Short Paper: Annotating a historical manuscript as a linguistic resource

H.-J. Döhla³, H. Klöter², M. Scholger¹, E. Steiner¹

1: University of Graz; 2: Humboldt-Universität zu Berlin; 3: Universität Tübingen

Keywords: Digital Scholarly Edition, Dictionary, Linguistics, Manuscript

Annotating a historical manuscript as a linguistic resource

The Bocabulario de lengua sangleya por las letraz de el A.B.C. is a historical Chinese-Spanish dictionary held by the British Library (Add ms. 25.317), probably written in 1617. It consists of 223 double-sided folios with about 1400 alphabetically arranged Hokkien Chinese lemmas in the Roman alphabet.

The contribution will introduce our considerations on how to extract and annotate linguistic data from the historical manuscript and the design of a digital scholarly edition (DSE) in order to answer research questions in the fields of linguistics, missionary linguistics and migration (Klöter/Döhla 2022).

Short Paper: How to Represent Topic Models in Digital Scholarly Editions

U. Henny-Krahmer¹, F. Neuber²

1: University of Rostock, Germany; 2: Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Keywords: text mining, topic modeling, digital scholarly editions, data modeling, data integration

How to Represent Topic Models in Digital Scholarly Editions

Topic modeling (Blei et al. 2003, Blei 2012) as a quantitative text analysis method is not part of the classic editing workflow as it stands for a way of working with text that in many respects contrasts with critical editing. However, for the purpose of a thematic classification of documents, topic modeling can be a useful enhancement to an editorial project. It has the potential to replace the cumbersome manual work that is needed to represent and structure large edition corpora thematically, as has been done for instance in the projects Alfred Escher Briefedition (Jung 2022), Jean Paul – Sämtliche Briefe digital (Miller et al. 2018) or the edition humboldt digital (Ette 2016).

We apply topic modeling to two edition corpora of correspondence of the German-language authors Jean Paul (1763-1825) and Uwe Johnson (1934-1984), compiled at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and the University of Rostock (Miller et al. 2018, Helbig et al. 2017). In our contribution, we discuss how the results of the topic modeling can be usefully integrated into digital editions. We propose to integrate them into the TEI corpora on three levels: (1) the topic model of a corpus, including the topic words and the parameters of its creation, is modeled as a taxonomy in a separate TEI file, (2) the relevance of the topics for individual documents is expressed in the text classification section of the TEI header of each document in the corpus, and (3) the assignment of individual words in a document to topics is expressed by links from word tokens to the corresponding topic in the taxonomy. Following a TEI encoding workflow as outlined above allows for developing digital editions that include topic modeling as an integral part of their user interface.

Short Paper: Analyzing the Catalogue of Heroines through Text Encoding

R. Milio

Bucknell University, United States of America

Keywords: Odyssey, heroines, prosopography, women

Analyzing the Catalogue of Heroines through Text Encoding

The Catalogue of Heroines (Odyssey 11.225-330) presents a corpus of prominent mythological women as Odysseus recounts the stories of each woman he encounters in the Underworld. I undertook a TEI close reading of the Catalogue in order to center ancient women in a discussion of the Odyssey and determine how the relationships between the heroines contribute to the Catalogue's overall purpose. In this short paper I demonstrate first my process: developing my own detailed feminist translation of the Catalogue, applying a TEI close reading to both my translation and the original ancient Greek, and creating a customized schema to best suit my purposes. Then, I detail my analysis of my close reading using cross-language encoding and a prosopography I developed through that reading, which reveals complex connections, both explicit and implied, among characters of the Catalogue. Third, I present the result of this analysis: that through this act of close reading I identified a heretofore unconsidered list of objects within the Catalogue and then demonstrated how these four objects of the Catalogue, ζώνη (girdle), βρόχος (noose), ἔδνα (bride-price), and χρυσός (gold), reveal the ancient Greek stigma surrounding women, sexuality, and fidelity. These objects clearly allude to negative perceptions of women in ancient Greek society and through these objects the Catalogue of Heroines reminds its audience of Odysseus' concerns regarding the faithfulness of his wife Penelope. Ultimately, by applying and adapting a TEI close reading, I identified patterns within the text that spoke to a greater purpose for the Catalogue and the Odyssey overall, that was able to export for further analysis of this prosopographical data. By the time of the conference, I will be able to present data visualizations that provide pathways that can assist other classicists to center women in ancient texts.

Session 7B – Long Papers – 09:30 - 11:00

Session 7B: Long Papers

Location: ARMB: 2.16

Chair: Gimena del Rio Riande, CONICET

Long Paper: Is it still data? Scholarly Editing of Text from Early Born-Digital Heritage

T. Roeder

Universität Würzburg, Germany

Keywords: TEI, born-digital heritage, retrocomputing, digitality, materiality

Dr Torsten Roeder

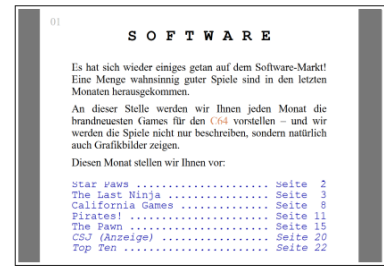
Zentrum für Philologie und Digitalität, Universität Würzburg (Germany)

Digital heritage is strongly bound to original devices and displays. Even in today's standardized environments, text can change its appearance depending on the monitor technology, on the processing software, and on the available fonts on the system: Text as data depends much on technical interpretation.

Creating a scholarly digital edition from born-digital heritage, especially text, needs to consider the original conditions, like encoding and hardware standards (cf. Roeder 2022). My question is: Are the encoding guidelines of the TEI suitable for representing born-digital text? How much information is required about the original environment? Can a screenshot serve as facsimile, or it is necessary to link to emulated states of the display software? (cf. Kaltman et al. 2021)

To give an example, I will present a preliminary scholarly TEI-based digital edition of “disk magazines” (one edition is currently available at https://diskmags.github.io/md_87-11.html). These magazines were a special type of periodical that was published mostly on floppy disk mainly in the 1980s and 1990s. Created by home computer enthusiasts for the community, disk magazines are potentially valuable as a historical resource to study the experiences of programmers, users and gamers in the early stage of microcomputing.

In the examples, the digital texts are decompressed byte sequences of PETSCII code, which is only partially compatible to ASCII. The appearance of the characters could be changed completely by the programmer to display foreign characters or alphabets. Further, it depended on a 40x25 characters layout, where text had to be aligned manually by inserting whitespaces. The once born-digital text – as data – is transformed into readable text – as image – on a screen. The example demonstrated that the connection between textual data and textual display can be very fragile.



Figures: Section “Software”, *Magic Disk 64*, no. 11, 1987, p. 1, in VICE Emulator (left), encoded in TEI-XML (middle), rendered HTML output in the browser (right).

For TEI encoding, this would have some consequences. On the one side, there would be a requirement to preserve as much of the original data as possible. On the other side, a scholarly edition needs to represent the semantics of the visible document. It would require an interpretative layer to communicate between these two levels, which could be implemented by different markup strategies; however it needs to be debated whether classes like “att.global.rendition” are actually suited for this (a discussion about this began last year on the TEI mailing list, cf. TEI-L 2021). It also needs to be discussed in which way a digital document (or which instance of it: as stored data, as memory state, as display?) can be interpreted in the same way as a material document – and which implications this would have for TEI encoding of born-digital heritage.

Bibliography

Kaltman, Eric; Osborn, Joseph; Wardrip-Fruin, Noah (2021). “From the Presupposition of Doom to the Manifestation of Code: Using Emulated Citation in the Study of Games and Cultural Software”. *Digital Humanities Quarterly* 15(1).

<http://www.digitalhumanities.org/dhq/vol/15/1/000501/000501.html>

Roeder, Torsten (2022): “Rescuing Diskmags: Towards Scholarly (Re-)Digitisation of an Early Born-Digital Heritage”, *Magazén* 3(1).

<https://edizionicafoscari.unive.it/it/edizioni4/riviste/magazen/>

TEI-L (2021). “Jakub Simek: “Rendition” encoding in born-digital documents”, *TEI-L Archives*, 19 October.

<https://listserv.brown.edu/cgi-bin/wa?A2=TEI-L;bbd3182.2110>.

Bibliography

The author currently works at the “Center for Philology and Digitality” at the University of Würzburg (Germany) in the department for digital scholarly editions. He is a member of the Institute for Documentology and Scholarly Editing. He studied musicology (PhD 2018) and Italian language and literature in Hamburg, Berlin and Rome. His research focuses on periodicals, retrocomputing, and textual variance.

Long Paper: Using Citation Structures

H. Cayless

Duke University, United States of America

Keywords: publishing, LOD, TEI infrastructure

Using Citation Structures

This paper is really a follow-up to one I gave at Balisage in 2021.[1] Citation Structures are a TEI feature introduced in version 4.2.0 of the Guidelines, which provide an alternative (and more easily machine-processable) method for declaring their internal structures.[2] This mechanism is important because of the heterogeneity of texts and consequently of the TEI structures used to model them. This heterogeneity necessarily means it is difficult for any system publishing collections of TEI editions to treat their presentation consistently. For example, a citation like “1.2” might mean “poem 1, line 2” in one edition, and “book 1, chapter 2” in another. It might be perfectly sensible to split an edition into chapters, or even small sections, for presentation online, but not at all to split a poem into lines (though maybe groups of lines might be desirable). A publication system otherwise will have to rely on assumptions and guesswork about the items in its purview, and may fail to cope with new material that does not behave as it expects. Worse, there is no guarantee that the internal structures of editions are consistent within themselves. We might consider, for example, Ovid’s ‘Tristia’, in which the primary organizational structure is book, poem, line, but book two is a single, long poem.

Citation structures permit a level of flexibility hard to manage otherwise, by allowing both nested structures and alternative structures at every level. In addition, a key new feature of citation structures over the older reference declaration methods is the ability to attach information that may be used by a processing system to each structural level. The <citeData> element which makes this possible will allow, for example, a structural level to indicate what name it should be given in a table of contents, or even whether or not it should appear in such a feature.

I will discuss the mechanics of creating and using citation structures. Finally, I will present a working system in XSLT that can exploit <citeStructure> declarations to produce tables of contents, split large documents into substructures for presentation on the web, and resolve canonical citations to parts of an edition.

Notes:

1. <https://www.balisage.net/Proceedings/vol26/html/Cayless01/BalisageVol26-Cayless01.html>
2. See <https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CORS6> and <https://tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SACRCS>.

Long Paper: Text between data and metadata: An examination of input types and usage of TEI encoded texts

T. Schaßan

Herzog August Bibliothek Wolfenbüttel, Germany

Keywords: Manuscript cataloguing, semantic markup, retro-conversion vs. born-digital

Text between data and metadata: An examination of input types and usage of TEI encoded texts

Many texts that have been encoded using the TEI in the past are retro-converted from printed sources: manuscript catalogues and dictionaries are examples for highly structured texts, drama, verse, and performance texts are usually less structured, editions appear somewhere inbetween.

Many of the text types for which the TEI offers specialised elements represent both metadata and data, according to the scenarios in which these texts are used.

In the field of manuscript cataloguing, it has been a question for a long time whether the msdescription module is sufficient for the representation of a retro-converted text of a formerly printed catalogue. One may argue, that a catalogue is first of all a visually structured text, a succession of paragraphs, whose semantics are only loosely connected to the main elements the TEI defines, such as <msContents>, <physDesc>, or <msPart>. On the other hand, on a sub-paragraph level, the TEI offers structures, which may not be align-able with the actual text of the catalogue so that the person who carries out the retro-conversion has to decide whether to change the text according to the TEI schema rules or encode the text semantically wrong or structure the text with much less semantic information as it would be possible.

Now, that the TEI is more and more used to store these kind of texts as born-digitals, the questions is whether the structures offered by the TEI meet all the needs the texts and their authors might have in different scenarios: Is a TEI-encoded text of a given kind equally useful for all search and computational uses, as well as publishing needs? Are the TEI structures flexible enough or do they privilege some uses over others? How much of the semantic information is encoded in the text and how much of it might be realised only in the processing of the sources?

In this paper, manuscript catalogues serve as an example for the more general question about what structures, how much markup and what kind of markup is needed in the time of powerful search engines and artificial intelligence, authority files and the Linked Open Data.

Session 8A – Long Papers – 11:30 - 13:00

Session 8A: Long Papers

Location: ARMB: 2.98

Chair: Meaghan Brown, Independent Scholar

Long Paper: Codex as Corpus : Using TEI to unlock a 14th-century collection of Old French short texts

S. Dows-Miller

University of Oxford, United Kingdom

Keywords: medieval studies; medieval literature; xforms; manuscript; codicology

Codex as Corpus : Using TEI to unlock a 14th-century Old French miscellany

Medieval manuscript collections of short texts can act as ready-made corpora, from which data may be drawn to help scholarship understand the circumstances of their production and early readership. As physical items they are of course materially discrete, yet this can obscure their internal diversity, containing stints by various scribes which often overlap asynchronously with texts from various exemplars, themselves from multiple geographical and temporal points of origin. The utility of traditional palaeographical and codicological techniques in examining such codices is well established, but there is also value to be found in the use of quantitative, digital methods that take the whole manuscript as the source of their data.

This paper will discuss the role played by TEI in an ongoing mixed-method study into one such codex-corpus: Bibliothèque nationale de France, fonds français, 24432, a fourteenth-century manuscript containing some 90 short texts written in Old French. The aim of the project has been to display how fruitful it can be to combine traditional and data-driven approaches in the holistic study of individual manuscripts, and its first phase has focussed on aspects of the manuscript's production, particularly through an examination of scribal hands.

TEI has been critical to this analysis, enabling discoveries about the manuscript which have eluded less technologically enabled generations of scholarship. For example, quantitative analysis of the rates and methods of scribal abbreviation, made possible through the manuscript's encoding, has supported and refined hypotheses reached through the qualitative analysis of palaeographic features regarding the number and division of hands within the manuscript. In addition to this confirmatory role, the quantitative analysis itself has led to further hypotheses which are less easily reached through qualitative methods, in particular in determining which variation is evidence of differences in scribal preferences, and which may be a continuation of variation found within the exemplar manuscripts.

As with any project of this nature, the process of encoding BnF fr. 24432 in TEI has not been without difficulty, and so this paper will also discuss the ways in which attempts have been made to streamline the process through automation and UI tools. These include the use of XForms in creating an input tool for the bare-bones encoding, and the use of XSLT workflows, both to perform tasks that would be overly time consuming if completed by hand, and also to create bespoke renders of the text encoded in TEI that are useful at a particular moment of the transcription or editing process. We will consider the benefits and drawbacks of creating one's own tools for such tasks, and the ways in which the tools created for this project may be usefully replicated or adapted by those conducting similar projects.

Long Paper: atop: another TEI ODD processor

S. Bauman¹, H. Bermúdez Sabel², M. Holmes³, D. Maus⁴

1: Northeastern University, United States of America; 2: University of Neuchâtel, Switzerland;

3: University of Victoria, Canada; 4: State and University Library Hamburg, Germany

Keywords: ODD, ODD chaining, RELAX NG, schema, XSLT Stylesheets

TEI is, among other things, a schema. That schema is written in and customized with the TEI schema language system “ODD”, for “one document does it all”. ODD is defined by Chapter 22 of the *Guidelines*, is used to *define* TEI P5, is used to *customize* TEI P5, and may also be used to define and customize non-TEI markup languages. The TEI supports a set of stylesheets (called, somewhat unimaginatively, “the Stylesheets”) that, among other things, convert ODD definitions of markup languages (including TEI P5) and customizations thereof into schema languages like RELAX NG and XSD that one can use to validate XML documents.

Holmes and Bauman have been fantasizing for years about entirely re-writing those Stylesheets, starting anew. Spurred by Maus' comment of 2021-03-23¹ Holmes presented a paper last year describing the problems with the current Stylesheets and, in essence, arguing that they should be re-written.² Within a few months the TEI Technical Council had charged Bauman with creating a Task Force for the purpose of creating, from scratch, an ODD processor that reads in one or more TEI ODD customization files, merges them with a TEI language (likely, but not necessarily, TEI P5 itself), and generates RELAX NG and Schematron schemas. It is worth noting that this is a distinctly narrower scope than the current Stylesheets,³ which, in theory, convert most any TEI into any of a variety of formats including DocBook, MS Word, OpenOffice Writer, Markdown, ePub, LaTeX, PDF, and XSL-FO (and half of those formats into TEI); and convert a TEI ODD customization file into RELAX NG, DTD, XML Schema, and ISO Schematron schemas, and into HTML documentation. A different group is working on the conversion of a customization ODD into customized documentation using TEIPublisher.⁴

The Task Force, which began meeting in April, comprises the authors. We meet weekly, with the intent of making slow, steady progress. Our main goals are that the deliverables be a utility that can be easily run on GNU/Linux, MacOS, or within oXygen, and that they be programs that can be easily maintained by any programmer knowledgeable about TEI ODD, XSLT, maybe XProc, and ant. Of course we also want the program to work properly. Thus we are generating test suites and performing unit testing (with XSpec⁵) as we go, rather than creating tests as an afterthought. We have also developed naming and other coding conventions for ourselves, and written constraints (mostly in Schematron) to help enforce them. So, e.g., all XSLT variables must start with the letter ‘v’, and all internal parameters must start with the letter ‘p’ or letters “tp” for tunnel parameters; all templates, variables, parameters, and function definitions must have their type explicitly declared with @as.

We are trying to tackle this enormous project in a sensible, piecemeal approach. We have (conceptually) completely separated the task of assembling one or more customization ODDs with a source ODD into a derived ODD from the task of converting the derived ODD into RELAX NG, and from converting the derived ODD into Schematron. In order to make testing-as-we-go easier, we are starting with the derived ODD → RELAX NG process, and expect to demonstrate some working code at the presentation.

The effort is open to public scrutiny at <https://github.com/TEIC/atop>.

Notes:

- 1 <https://github.com/TEIC/Stylesheets/pull/477#issuecomment-805138553>
- 2 “Are the TEI Stylesheets really broken beyond repair?” (Conference presentation.) Text Encoding Initiative Conference 2021 (online). 2021-10-25.
- 3 <https://github.com/TEIC/Stylesheets/>
- 4 <http://teipublisher.com/exist/apps/tei-publisher-home/index.html>
- 5 <https://www.xml.com/articles/2017/03/15/what-xspec/>

Session 8B – Demonstrations – 11:30 - 13:00

Session 8B: Demonstrations

Location: ARMB: 2.16

Chair: Tiago Sousa Garcia, Newcastle University

Demonstration: Transcribing Primary Sources using FairCopy and IIIF

N. Lailaona

Performant Software Solutions LLC, United States of America

Keywords: Digital Humanities Critical Editions Tools IIIF

Transcribing Primary Sources using FairCopy and IIIF

FairCopy is a simple and powerful tool for reading, transcribing, and encoding primary sources using the TEI Guidelines. FairCopy can import IIIF manifests as a starting point for transcription. Users can then highlight zones on each surface and link them to the transcription. FairCopy exports valid TEI-XML which is linked back to the original IIIF endpoints. In this demonstration, we will demonstrate the IIIF functionality in FairCopy and then take a look at the exported TEI-XML and how it provides a consistent interface to images as well as the original IIIF manifest.

Demonstration: Adapting CETEIcean for static site building with React and Gatsby

R. Viglianti

University of Maryland, United States of America

Keywords: Digital publishing, TEI processing, static sites, programming

Adapting CETEIcean for static site building with React and Gatsby

Raffaele Viglianti, Maryland Institute for Technology in the Humanities, University of Maryland

The JavaScript library CETEIcean, written by Hugh Cayless and Raff Viglianti, relies on the DOM processing of web browsers and HTML5 Custom Elements to publish TEI documents as a component pluggable into any HTML structure. This makes it possible to publish and lightly transform TEI documents directly in the user's browser, doing away with complex server-side infrastructure for TEI publishing. This lightweight approach to publishing can be valuable in a "text as data" context, where the focus of labor and algorithmic complexity may be more centered on corpus building and analysis as opposed to publication. However, CETEIcean provides a fairly bare-bones API for a fully-fledged TEI publishing solution and, without some additional considerations, TEI documents rendered with CETEIcean can be invisible to search engines.

This demonstration will showcase an adaptation of the CETEIcean algorithm as a plugin for the static site generator Gatsby, which relies on the popular framework React for building user interfaces (UI). The static site pages generated with Gatsby will contain embedded TEI data, making it visible to search engines. Two plugins will be shown:

- gatsby-transformer-ceteicean (<https://www.gatsbyjs.com/plugins/gatsby-transformer-ceteicean/>) prepares XML to be registered as HTML5 Custom Elements. It also allows users to apply custom transformations before and after processing if the TEI data requires it for publication (the demonstration will show an example related to addSpan elements).
- gatsby-theme-ceteicean (<https://www.npmjs.com/package/gatsby-theme-ceteicean>) implements HTML5 Custom Elements for XML publishing, particularly with TEI. It re-implements parts of CETEIcean excluding behaviors; instead, users can define React components to customize the behavior of specific TEI elements. This makes it possible to access powerful React functionalities such as state management for user interaction.

The demonstration will show examples from the *Scholarly Editing* journal (<https://scholarlyediting.org>), which published TEI-based small-scale editions with these tools alongside other essay-like content.

Biography:

Dr. Raffaele (Raff) Viglianti is a Senior Research Software Developer at the Maryland Institute for Technology in the Humanities, University of Maryland. His research is grounded in digital humanities and textual scholarship, where “text” includes musical notation. He researches new and efficient practices to model and publish textual sources as innovative and sustainable digital scholarly resources. Dr. Viglianti is currently an elected member of the Text Encoding Initiative technical council and the Technical Editor of the *Scholarly Editing* journal.

Demonstration: Spec Translator: Enabling translation of TEI Specifications

H. Cayless

Duke University, United States of America

Keywords: TEI, Translation, crowdsourcing

Spec Translator: Enabling translation of TEI Specifications

This demonstration will introduce Spec Translator, available from <https://translate.tei-c.org/> which enables users to submit pull requests for translations of specification pages from the TEI Guidelines.

Demonstration: LEAF-Writer: a TEI + RDF online XML editor

D. Jakacki¹, S. Brown², J. Cummings³

1: Bucknell University, United States of America; 2: University of Guelph, Canada; 3: Newcastle University, UK

Keywords: TEI, RDF, Online Editors

LEAF-Writer: a TEI + RDF online XML editor

LEAF-Writer is an open-source, open-access Extensible Markup Language (XML) editor that runs in a web browser and offers scholars and students a rich textual editing experience without the need to download, install, and configure proprietary software, pay ongoing subscription fees, or learn complex coding languages. This user-friendly editing environment incorporates Text Encoding Initiative (TEI) and Resource Description Framework (RDF) standards, meaning that texts edited in LEAF-Writer are interoperable with other texts produced by the scholarly editing community and with other materials produced for the Semantic Web. LEAF-Writer is particularly valuable for pedagogical purposes, allowing instructors to teach students best practices for encoding texts without also having to teach students how to code in XML directly. LEAF-Writer is designed to help bridge the gap by providing access to all who want to engage in new and important forms of textual production, analysis, and discovery. LEAF-Writer draws on TEI All as well as other TEI-C-supplied schemas, can use project-specific customized schemas, and offers continuous validation against supported and declared schemas. LEAF-Writer allows users to access and synchronize their documents in GitHub and GitLab, as well as to upload and save documents from their desktop. This presentation will demonstrate the variety of functionality and affordances of LEAF-Writer.

Virtual Poster Session on Gather.Town – Thursday 22 September 2022

Virtual Poster Session on Gather.Town

Virtual location: Gather.Town

Chair: Martina Scholger, University of Graz

A Virtual Poster session will be held in <https://gather.town/> on the Thursday after the conference (22 September 2022) to enable people to participate who are not able to physically attend the conference. Accepted poster presenters from the conference will automatically be eligible to present in the Virtual Poster session as well.

The Virtual Poster session will be held in <https://gather.town/> on the Thursday after the conference, **22 September 2022 at 1:00pm BST**. All physical poster presenters or virtual posters accepted to the conference are able to present in the Virtual Poster session. The poster session will be run using <https://gather.town/> where we set up a space for each poster presenter: <https://app.gather.town/app/DVLCOOcP1ITL5Zkh/TEI2022> (the link is only available during the poster session).

To participate, please send your poster to martina.scholger@uni-graz.at by **Tuesday 20 September 2022** at the latest, taking into account the following specifications:

- PNG or JPG format
- landscape format
- min. 1000 x 600 pixels
- max. 3 MB

It is expected that at least one of the authors will be available to answer questions during the poster session.

The presentations submitted solely for virtual posters are listed below.

Virtual Poster: From Archives to TEI Publisher: Digital Edition of German Work Regulations in the Project 'Non-state Law of the Economy'

P. Solonets

Max Planck Institute for Legal History and Legal Theory, Germany

Keywords: digital edition, text processing, data management, tei publisher, ocr

From Archives to TEI Publisher: Digital Edition of German Work Regulations in the Project 'Non-state Law of the Economy'

Abstract

The aim of the project 'Non-state Law of the Economy' is to build a digital collection of primary sources, showing the normative world of industrial relations in the German metal industry of the 19th-20th centuries. This collection includes various types of textual documents from collective agreements to company pension insurances. This poster's focus is on the life cycle of textual data inside the project from archives to TEI Publisher.

The project makes wide use of computer-assisted processing and DH-tools. The workflow begins in archives, where the textual data is collected using smartphones and a scan tent. As a next step, an open-source OCR software is applied to the scans to produce output in form of XML Pages, which subsequently transformed to a standard TEI XML. After a basic structural annotation and correction, the texts are uploaded to TEI Publisher.

For scholarly text analysis, our team developed a key word tree serving as the basis for the classification of legal and social relations and norms present in the texts. This classification is applied in the form of annotation, with certain terms highlighted and marked up as belonging to one of the categories from our key word tree. The annotation takes place online in our own instance of TEI Publisher. Currently we are working on the automation of the annotation process.

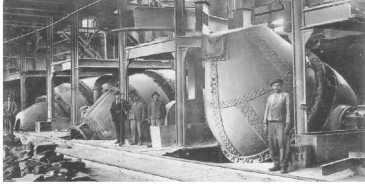
Our texts are open for the interested scholarly community. They are available for download in different formats (including TEI XML) and we expect texts to be reused by scholars for their own research.

Now there are around 60 sources published at our instance of TEI Publisher with more 300 on the way. We aim at improving the handling of our data, introducing more automation to the data editing at all stages and broadening the functionality of our instance of TEI Publisher.

Biography:

Polina Solonets works as a researcher at Max Planck Institute for Legal History and Legal Theory, collaborating on several projects there and facilitating the implementation of digital tools and methods. She is involved in the preparation of scholarly digital editions in the joint project 'The School of Salamanca' and the research project 'Non-State Law of the Economy'. She also takes part in the project 'HyperAzpilcueta. Visualising the instability of early modern normative knowledge'.

From Archives to TEI Publisher: Digital Edition of German Work Regulations in the Project 'Non-state Law of the Economy'



About the Project

The aim of the project is to build a digital collection of primary sources, showing the normative world of industrial relations in the German metal industry of the 19th-20th centuries

Sources

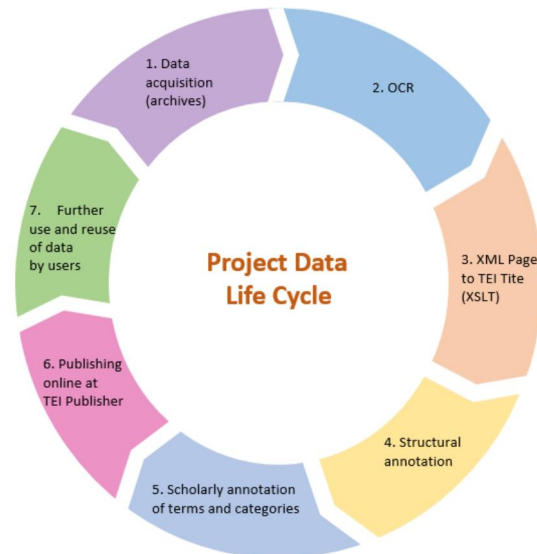
various types of textual documents such as collective bargaining agreements, company pension insurances, internal work regulations, company flats rental contracts, apprenticeship contracts, etc.

Technology

OCR, XSLT, TEI XML, Python, TEI Publisher

Current Results

- 60 sources published so far;
- 300 more are on their way;
- our instance of TEI Publisher is being constantly developed and broadened in its functionality
- planning to improve the handling of our data and introduce more automation



Polina Solonets (solonets@ihl.mpg.de)



MAX PLANCK INSTITUTE
FOR LEGAL HISTORY
AND LEGAL THEORY

Virtual Poster: Feature structures for character social variable annotation and an application to Alsatian theater

P. Ruiz Fabo¹, H. Bermúdez Sabel²

1: Université de Strasbourg, France; 2: Université de Neuchâtel, Switzerland

Keywords: feature structures, character analysis, theater, personography, Alsatian

Feature structures for character social variable annotation and an application to Alsatian theater

Pablo Ruiz Fabo (Université de Strasbourg), Helena Bermúdez Sabel (Université de Neuchâtel)
Several works address the computational treatment of dramatic characters. Zöllner-Weber (2008, 2011) presents a character analysis ontology. Galleron (2017) developed a *characteriseme* (characterization unit) taxonomy based on character lists in French theater between 1630 and 1810, formalized as a TEI feature structure (FS) library (see Romary, 2015). Following Phelan (1989), the taxonomy includes *mimetic* features, which give characters traits assimilating them to humans, and synthetic ones, describing their role in the plot.

We believe that *characteriseme* analysis using a common annotation schema can help comparative drama analysis. We successfully adapted Galleron's FS approach to model characters in a different language (Alsatian) and period (1870-1940). This can help compare the Alsatian tradition to the hegemonic literatures surrounding it (German and French), one of the goals towards which our ongoing MeThAL project contributes (Ruiz et al., 2022).¹

The poster's contributions:

- A character feature (*characteriseme*) taxonomy using feature structures, inspired by Galleron (2017) but providing an improved, more modular implementation, and enabling the description of more recent drama
- A TEI personography where each of our corpus' 2386 characters is described according to the feature structure²
- First characterization analyses in the corpus based on it

Intermediate levels were added in our FS to better group *mimetic* features into basic traits (age, gender, origin, language), socioeconomic traits (profession, class) and relation-position traits (where a character stands in personal or professional relations, e.g. *spouse* or *manager*). Controlled vocabularies were added, including a list of ca. 350 professions and a taxonomy of socioprofessional groups. Personography compliance was ensured with a schema automatically derived from the FS System Declaration (Bermúdez, 2019). The annotations have yielded insight into how female characters are characterized differently by female authors (increased reference to character's profession) vs. male ones. An interface to navigate the corpus based on the annotations was created.³

References

Bermúdez Sabel, H. (2019). Encoding of Variant Taxonomies in TEI. *Journal of the Text Encoding Initiative*, Issue 11, Article Issue 11. <https://doi.org/10.4000/jtei.2676>

Galleron, I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions. *Human and Social Studies*, 6(1), 88–108.

<https://doi.org/10.1515/hssr-2017-0007>

Phelan, J. (1989). *Reading people, reading plots: Character, progression, and the interpretation of narrative*. University of Chicago Press.

Romary, L. (2015). Standards for language resources in ISO – Looking back at 13 fruitful years. *Edition - Die Fachzeitschrift Für Terminologie*, 11(2), 13–19.

<https://hal.inria.fr/hal-01220925>

Ruiz Fabo, P., Bernhard, D., & Werner, C. (2022). The benefits of increasing the digital availability of Alsatian theater. *Digital Humanities 2022*, 567–560.

<https://doi.org/10.5281/zenodo.7014965>

Zöllner-Weber, A. (2008). *Noctua literaria: A computer-aided approach for the formal description of literary characters using an ontology* [Universität Bielefeld].

<https://core.ac.uk/reader/15958020>

Zöllner-Weber, A. (2011). Text encoding and ontology—Enlarging an ontology by semi-automatic generated instances. *Literary and Linguistic Computing*, 26(3), 365–370.

<https://doi.org/10.1093/lc/fqr021>

Appendix: Two personography entries based on the FS library

```
<person xml:id="mtl-per-0890">
<bibl corresp="#mtl-090"/>
<persName>Alice Sandel</persName>
<note type="roleDesc">Dactylo</note>

<occupation>Dactylo</occupation>
<fs type="character_specification">
<f name="specification_type">
<fs type="mimetic_features">
<f name="general">
<fs type="general_features">
<f name="sex">
<symbol value="F"/>
</f>
</fs>
</f>
<f name="socio_economic_status">
<vColl>
<fs type="professional_activities">
<f name="occupation">
<symbol value="typist"/>
</f>
<f name="professional_category">
```

```
<person xml:id="mtl-per-0328">
<bibl corresp="#mtl-031"/>
<persName>Dr. Schröpfer</persName>
<note type="roleDesc">Doktor, älterer Herr,
weiße Perücke, eleganter Anzug</note>

<occupation>Doktor</occupation>
<fs type="character_specification">
<f name="specification_type">
<fs type="mimetic_features">
<f name="general">
<fs type="general_features">
<f name="sex">
<symbol value="M"/>
</f>
</fs>
</f>
<f name="socio_economic_status">
<vColl>
<fs type="professional_activities">
<f name="occupation">
<symbol value="doctor"/>
</f>
<f name="professional_category">
```


<pre> <symbol value="intermediate_professionals"/> </f> </fs> <fs type="socio_economic_other"> <f name="social_class"> <symbol value="lower_class"/> </f> </fs> </vColl> </f> <f name="language"> <default/> </f> </fs> </f> </fs> </person> </pre>	<pre> <symbol value="professionals_scientific_technical"/> </f> </fs> <fs type="socio_economic_other"> <f name="social_class"> <symbol value="upper_class"/> </f> </fs> </vColl> </f> <f name="language"> <default/> </f> </fs> </f> </fs> </person> </pre>
---	---

Fig. 1: Two examples of characters described through FS. Besides the FS, a <note type="roleDesc"> element provides the original information in the *dramatis personæ* based on which the feature values were established.

Notes:

1 <https://methal.pages.unistra.fr/en>

2 See <https://git.unistra.fr/methal/methal-sources/-/tree/master/personography> for the FS library, personography and schema

3 <https://methal.eu/ui/>

Feature structures for character social variable annotation and an application to Alsatian theater

Pablo Ruiz Fabo¹, Helena Bermúdez Sabel² · ¹Université de Strasbourg, ²Université de Neuchâtel

Background

Alsatian theater and its analysis

- Alsatian? Germanic varieties from Alsace (Eastern France)
- Rich dramatic tradition, since early 19th century
- Genres
 - Comedy predominates (farces to refined satire)
 - Popular dramas (*Volksstücke*) + tales (*Måreli*)
- Language varieties (standard vs. dialect) used as characterization means
- No previous large-scale quantitative analyses

MeThAL project

- Towards a macroanalysis of theater in Alsatian
- First Alsatian theater electronic corpus (TEI)
- > 500,000 tokens (in progress), 51 plays
- First resources for quantitative analysis
- Publicly available:
 - TEI plays, TEI personography, emotion lexicon, user interface

Character social variable annotation

- Links with earlier studies of the tradition (focused on the plays' social makeup)
- *Dramatis personæ* hints at plots and conflicts
- Feasible while entire text still under encoding

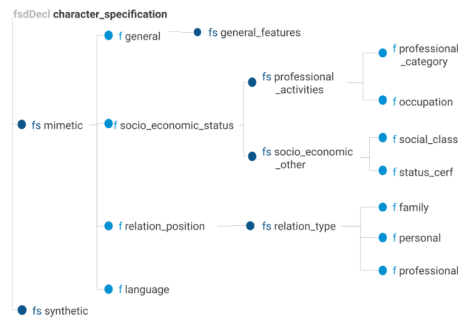
Socioprofessional groups

- professionals, scientific, technical
- intermediate professions
- service and sales
- crafts
- industry and transportation
- agriculture
- elementary professions
- renters
- clergy
- military
- government officials
- associative world

References

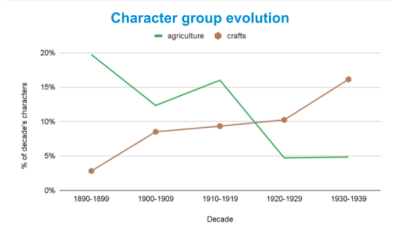
Bermúdez Sabel, H. (2020). Encoding of Variant Taxonomies in TEI. *Journal of the Text Encoding Initiative*, 1(1).
 Galleron, I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions. *Human and Social Studies*, 6(1), 88-108.
 Phelans, J. (1989). *Rising people, reading plots: Character progression and the interpretation of narrative*.
 Romary, L. (2015). Standards for language resources in ISO - Looking back at 13 fruitful years. *Edition - Die Fachzeitschrift für Terminologie*, 11(2), 13-19.

Feature library: Taxonomy of characterisemes (characterization units)



- Phelans (1989) mimetic characterisemes
- Formalized with TEI Feature Structures
- Improvement over Galleron (2017), adding modularity
- Personography with 2,386 characters (232 plays)

Findings



Characterization

Based on the <i>dramatis personæ</i>	Female authors	Male authors
Characters with a profession	35.71%	48%
- Female characters among these	45.71%	11.3%
- Male characters among these	54.29%	85.43%

Structured corpus navigation

Professional Group Social Class Character gender 14 plays | 5 authors

Character cooccurrences in cast list, by professional group

Characters: 98 (7: 72, M: 26) | With prof group: 55 (35.71%; F: 14 - 45.71%, M: 19 - 54.29%)

	agri	craf	elem	gov	inte	pro	seSa
agri	11.46	0	3.28	3.28	6.56	0	15.11
craf	0	9.84	9.84	3.28	0	1.64	0
elem	3.28	9.84	0	4.92	0	1.64	3.28
gov	3.28	3.28	4.92	1.64	0	3.28	0

ODD automatically derived from fsdDecl with tool in Bermúdez (2020)

Virtual Poster: Multilingualism and multiscrptism in TEI publishing: DH2022

Y. Wang^{1,2}, K. Nagasaki¹, I. Ohmukai³, M. Shimoda³

1: International Institute for Digital Humanities, Japan; 2: Graduate school of the University of Tokyo; 3: The University of Tokyo

Keywords: language, script, typography, multilingualization

Multilingualism and multiscrptism in TEI publishing: DH2022

Yifan Wang, Kiyionori Nagasaki, Ikki Ohmukai, Masahiro Shimoda

In the conference DH2022 Tokyo held this July, the book of abstracts has been published entirely through the XSL-FO pipeline based on ADHO's DHConvalidator and TEI to PDF Book Creator. The texts of the abstracts were converted by each author with DHConvalidator, which generates a format not always expected by the original TEI to PDF Book Creator. Moreover, while the text body is mostly written in English and some other languages which are accepted in CFP, it also embraces a large number of words and phrases in various Asian languages, reflecting the theme of the conference and authors' regional backgrounds. Thus, we needed to adapt the original stylesheets to multi-script typography by a large expansion of linguistic and typographical templates as well as extensive annotation.

Our modification involves extraction and annotation of Asian language fragments in TEI documents, locale-oriented typeface differentiation, adjustment for typographical conventions, and mixed script typesetting. We will share our methodology and decisions we applied to the actual book, hoping that it serves as a case study that leads to dissemination of attention to, and better practices in, non-Latin and/or multi-script publication in the TEI community.

Multilingualism and multiscrptism in TEI publishing: DH2022

Yifan Wang (International Institute for Digital Humanities / Graduate school of the University of Tokyo), Kiyionori Nagasaki (International Institute for Digital Humanities), Ikki Ohmukai (The University of Tokyo), Masahiro Shimoda (Graduate school of the University of Tokyo)

Language-based markup	Locale-specific typefaces	Mixed-script typesetting		
<p>Report, Vol. 33, pp. 1-11.</p> <p>[7] Baochang, G. 耿宝昌 (1993). <i>Ming Qi jianming mingqi wuzhi yanjiu</i> [Ming and Qing Porcel Inspection]. Beijing: The Palace Museum.</p> <p>[8] Jun Z. 朱军 (2002). <i>Mingqi Qingchu huanyu jianming</i> 明代清初青花瓷器鉴定 [A la early Qing dynasty blue and white gobllet ident]. <i>Wenwu Shijie</i> 文物世界 4, pp. 38-42.</p> <p>[9] Jingjing, X. 徐菁菁 (2017). <i>Mingqing yuanyu yi tezheng</i> 明清瓷器图案及特征 [Sous Characteristics of Ming and Qing beaker vases]. <i>艺术品</i>, 11, pp. 66-73.</p> <p>[10] Medley, M. (1987). "The Ming-Qing Chinese Porcelain", <i>Artis Antiquar</i> 42, pp. 65.</p> <p>[11] Jenyas, S. (1955). "The Wares of the T. Period Between the Ming and Ch'ing, 1620-16</p> <p>We have published the book of abstracts of DH2022 Tokyo using XSL-FO software suite provided by ADHO. The abstracts contain a large number of names and phrases in multiple Asian languages, which are not supported by the default setup. We thus decided to make a large expansion on its multilingual function.</p> <pre></bibli> <tbl_struct> <tbl_header> <tr><td>Vann, Yipso</td><td>683</td></tr> <tbl_info cols="2"> <tbl_r cells="2" ix="1" maxcspan="1" maxrspan="1" usedcols="2"><\/tbl_r> <tbl_r cells="2" ix="2" maxcspan="1" maxrspan="1" usedcols="2"><\/tbl_r> <tbl_r cells="2" ix="3" maxcspan="1" maxrspan="1" usedcols="2"><\/tbl_r> <tbl_r cells="2" ix="4" maxcspan="1" maxrspan="1" usedcols="2"><\/tbl_r> <tbl_r cells="2" ix="5" maxcspan="1" maxrspan="1" usedcols="2"><\/tbl_r> </pre> <p>Author index is improved to accept tailored name sorting and normalization of spelling variants based on manual annotation in TEI files.</p> <p>☺ Ideally, the language of the text should be specified in @lang in TEI, but we added them via custom @rend values following the existing style processing logic of TEI to PDF Book Creator.</p>	Vann, Yipso	683	<p>返 返</p> <p>返 返</p> <p>There are also passages of medieval Latin abbreviation, Sanskrit, and Thai found in the source files, that we needed to find specific fonts for. For the complex layout scripts, some fonts seemed to have minor incompatibilities with the processor (Apache FOP), and we needed to select a working font for our environment.</p> <p>☺ CJK languages are accustomed to use different fonts for each locale, but most Cyrillic fonts implement regional variants through the fonts' OpenType features. That would make localization difficult in XSL-FO processing with Apache FOP, which currently does not support OpenType features.</p>	<p>Intertextuality has b scholarly communities : like <i>Redaktionsgeschic</i> 校助學 in China have I foundations for debates works, editions and autl</p> <p>Intertextuality has b scholarly communities : like <i>Redaktionsgeschic</i> 校助學 in China have I foundations for debates works, editions and autl</p> <p>Edge case: some Thai typefaces are designed smaller to allow the safe margin for stacking diacritics. Thai letters were enlarged lest be awkwardly undersized among ordinary Roman typefaces.</p> <p>☺ Although our abstracts contain small portions of Hebrew, RTL scripts did not cause us problems as far as we know, mostly because the default Times New Roman font supports Hebrew out of the box. We did not test how markup-based font specification interacts with RTL intermixing.</p> <p>Our modified version of TEI to PDF Book Creator program for DH2022 is available at: https://github.com/747/tei-to-pdf-dh2022/</p>
Vann, Yipso	683			

Virtual Poster: Celebrating Deviation: Encoding Variant Japanese Phonetic Characters known as Hentaigana

H. McGaughey

Hosei University, Japan

Keywords: Japanese script, East Asian texts, character variation, hentaigana, digital editions

Celebrating Deviation: Introducing Markup Options for Variant Japanese Phonetic Characters known as *Hentaigana*

In digitalizing the manuscript heritage of secret writings by the Japanese 15th century actor, playwright, producer, and teacher Zeami, I am including the premodern script variations known as *hentaigana* now available in Unicode. *Hentaigana* are variant *hiragana*, phonetic characters that are used to write various Japanese grammatical and function words and often ruby, for which the TEI released elements last year. In 2017, Unicode formally added 285 *hentaigana* characters in their Kana Supplement and Kana Extended-A code charts. These alternatives are fluid or cursive abbreviations of various “parent characters,” phonetically used Chinese characters (*kanji*), with varying patterns and degrees of simplification. In encoding both the modern, standardized hiragana and *hentaigana*, this project makes the manuscripts more accessible to learners of Japanese premodern script. Comparisons of the variants in different text witnesses using such markup might be useful for future analyses of text genealogy.

In this poster, I will present my methods for systematically including *hentaigana* developed while transcribing manuscript witnesses of Zeami’s writings and explain my inclusion of the “old character forms” (*kyūjitai*) of *kanji* using similar markup. I will furthermore share initial orthographical explorations of texts encoded thus far and consider methods for sharing the project with digitally savvy users, noh theater experts with no IT background, and educated non-specialists.

Celebrating Deviation TEI

Introducing Markup Options for Variant Japanese Phonetic Characters known as *Hentaigana*
 Hanna McGaughey

Intro

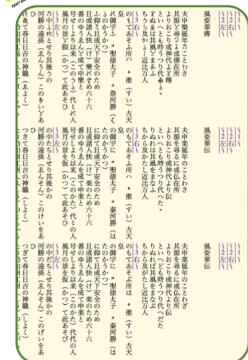
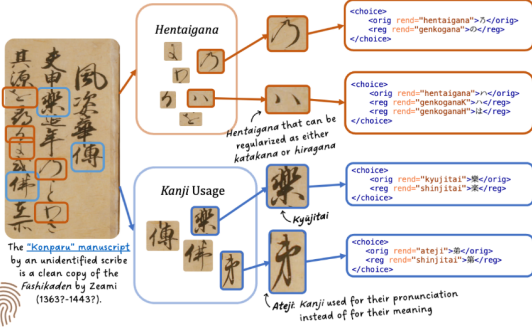
What are hentaigana?
 Ex. with characters representing /a/
 literally "variant" or "deviant kana"
 1900 script Reform → あ
 Simplifications derived from different "parent" kanji (Sinoglyphs)...
 ...and various degrees of simplification

What are kyūjitai?
 literally "old character forms"
 學 学 → 学 "to study"
 幸 幸 → 幸 "other reforms"
 聲 声 → 声 "voice"
 Complex premodern kanji → Modern standardized simplifications

Why include this markup?

- To improve the accessibility of premodern manuscripts
- To investigate scribes' unique handwriting "fingerprints"

Workflow

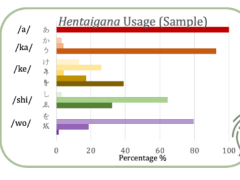
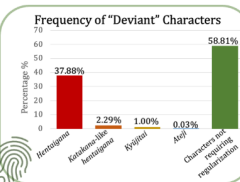


Prototype for a digital scholarly edition (DSE) to improve manuscript accessibility

What needs to be done next?

- Finish transcribing the "Konparu" manuscript
- Transcribe other manuscripts to fill in gaps, etc., in the DSE and to compare "fingerprints"
- Develop more "fingerprint" statistics about locations of specific hentaigana (ex. へ for particles), inconsistencies in character usage (ex. 備古, さいこ, and さいこ for keiko), etc.

The Unidentified Scribe's Handwriting "Fingerprint:"



Preliminary Results

野上記念法政大学能楽研究所
 The Nogami Memorial Noh Theatre Research Institute of Hosei University

法政大学
 HOSEI UNIVERSITY

日本学術振興会
 Alexander von Humboldt Stiftung/Foundation

My sincere thanks to Prof. Yamanaka Reiko and Yi Suyang for their support and contribution, and to Luka and Daniel Slane for their love and understanding.

Thanks

Virtual Poster: Theoretical and practical challenges of automatically identifying and encoding alliteration in texts written in Italian

A. Consalvi¹, S. Fumagalli²

1: Sapienza University of Rome; 2: Università Cattolica del Sacro Cuore, Milan

Keywords: alliteration, XML-TEI, encoding, poetry, translation

Theoretical and practical challenges of automatically identifying and encoding alliteration in texts written in Italian

In our proposed presentation, we would like to display the theoretical and practical challenges posed by the creation of a program aimed at automatically identifying and encoding alliteration in texts written in Italian. Furthermore, a reflection on the possibilities offered by the analysis of the abovementioned phenomenon will be presented: from examining the style of a poet to determining if and how this literary device is preserved in translation.

On the theoretical level, alliteration is generally defined as a literary device consisting of the repetition of sounds at the beginning of adjacent words (cf. Beltrami, 2011). But what kind of sounds are we talking about? How long should they be? And what do we mean by ‘adjacent’? These are all crucial interrogatives that must be dealt with at the very beginning of any investigation on alliteration, especially considering that scholars (cf. Valesio, 1967; Lausberg, 1969; Menichetti, 1993; Mortara Garavelli, 1997; Ellero and Redisori, 2001; Ghiazza and Napoli, 2007; Mortara Garavelli, 2010; Arduini and Damiani, 2010; Lavezzi, 2017; Motta, 2020) tend to offer slightly different definitions.


On the practical level, once the rule-based program is created for Italian, it can be easily adapted to languages with a high degree of correspondence between graphemes and phonemes. Given a TXT file, the program is likely to be able to automatically identify the above-mentioned phenomenon. However, in this case the demanding task is the encoding process: a thorough reflection is needed to find a proper way to define an XML-TEI tag that contains all the important information such as the repeated sound and the number of words involved.

- Valesio, P., (1967). *Strutture dell'allitterazione: grammatica, retorica e folklore verbale*. Bologna: Zanichelli.
- Lausberg, H., (1969). *Elementi di retorica*. Bologna: Il mulino.
- Menichetti, A., (1993). *Metrica italiana: fondamenti metrici, prosodia, rima*. Padova: Antenore.
- Mortara Garavelli, B., (1997). *Manuale di retorica*. Milano: Bompiani.
- Ellero, M. P. and Residori, M., (2001). *Breve manuale di retorica*. Milano: Sansoni.
- Ghiazza S. and Napoli, M., (2007). *Le figure retoriche: parola e immagine*. Bologna: Zanichelli.
- Mortara Garavelli, B., (2010). *Il parlar figurato. Manuale di figure retoriche*. Bari: Laterza.
- Arduini, S. and Damiani, M., (2010). *Dizionario di retorica*. LacCom.
- Beltrami, P. G., (2011). *La metrica italiana*. Bologna: Il mulino.

- Lavezzi, G., (2017). *Breve dizionario di retorica e stilistica*. Roma: Carocci.
- Motta, U., (2020). *Lingua mortal non dice: guida alla lettura del testo poetico*. Roma: Carocci.

Andrea Consalvi obtained his BA (2016) and MA (2018) degrees in Foreign Language and Literature at Università Cattolica del Sacro Cuore (Brescia) and earned his PhD degree (2022) in Linguistics and Foreign Languages from the Milan campus of UCSC. His research interests include digital humanities, authorship attribution and linguistics. At Sapienza University of Rome, he worked as a postdoc (2019–2020) for the PRIN ‘The Transmission of Ancient Linguistics: Texts and Contexts of the Roman Grammatical Studies’ and he is now working for the ERC ‘Priscian's Ars Grammatica in European Scriptoria’.


Stefano Fumagalli obtained his BA (2018) and MA (2020) degrees in Foreign Language and Literature at Università Cattolica del Sacro Cuore (Milan) where he is currently a PhD candidate in Linguistics and Foreign Languages. His research interests include Russian literature, metre and prosody, and poetic translation.



Theoretical and practical challenges of automatically identifying and encoding alliteration in texts written in Italian

Andrea Consalvi¹ and Stefano Fumagalli²

¹Sapienza University of Rome
²Università Cattolica del Sacro Cuore



SOURCE	NATURE OF SOUND(S)	POSITIONS OF SOUND(S)	DISTANCE BETWEEN WORDS	No. OF WORDS
Valesio 1967	phoneme or allophones of the same phoneme	beginning or ending	/	three or more
Lausberg 1969	consonant or syllable beginning with a consonant	beginning	/	/
Menichetti 1993	mostly consonant sounds	originally at the beginning	/	/
Garavelli 1997	consonant or syllable	beginning	consecutive	/
Ellero-Residori 2001 (strict def.)	phoneme	beginning	"consecutive words, or at least at close distance, belonging to the same textual segment"	mostly two words
Ellero-Residori 2001 (extended def.)	any sound	beginning or within	words at close distance	mostly two words
Ghiazza Napoli 2007	sound (consonant or vowel)	beginning or within	close words	/
Garavelli 2010	vowels, consonants or syllables	beginning or within (or partly at the beginning and partly at the end)	consecutive	two or more
Arduini-Damiani 2010	consonant or syllable	/	/	/
Beltrami 2011	sounds or groups of sounds	beginning	consecutive or at close distance	/
Lavezzi 2017	consonant (or syllable)	beginning (or within)	consecutive or at close distance	/
Motta 2020	similar or identical sounds (consonants or vowels)	beginning (or within)	/	two or more

Research question

Is it possible to code a rule-based program to automatically identify and encode alliteration? Maybe, but when it comes to this literary device several definitions have been given and most of them are slightly different. Therefore, a theoretical reflection is needed to understand whether the rules of alliteration can be listed unambiguously.

Encoding alliteration

```

<?xml version="1.0" encoding="UTF-8" >
<text>
  <body>
    <lg>
      <l>Ma ben veggio or si come al popol tutto</l>
      <l>
        <s>
          <w xml:id="a01">favola</w>
          <w xml:id="a02">fui</w>
        </s>
      </l>
      <l>gran tempo, onde sovente
      <l>di
        <s>
          <w xml:id="a03">me</w>
          <w xml:id="a04">medesmo</w>
          <w xml:id="a05">meco</w>
          <w xml:id="a06">mi</w>
        </s>
      </l>
      <l>vergogno;
    </lg>
    <spanGrp>
      <span type="alliteration" n="2" from="#a01" to="#a02"></span>
      <span type="alliteration" n="3" from="#a03" to="#a05"></span>
      <span type="alliteration" n="4" from="#a03" to="#a06"></span>
    </spanGrp>
  </body>
</text>
  
```

Research possibilities

O Tite tute Tati tibi tanta tyranne tulisti

■ O Tito Tazio tiranno, tu stesso ti attirasti atrocità tanto tremende!

■ Thou tyrant, Titus Tatius, such great troubles you brought upon yourself

■ O Тит Татий тиран, тяготал тебя тяготы эти!

■ Oh Tito Tacio, tirano, tú mismo te produjiste tan terribles desgracias!

Bibliography

- P. Valesio, *Strutture dell'allitterazione: grammatica, retorica e folklore verbale*, Zanichelli, Bologna 1967.
- H. Lausberg, *Elementi di retorica*, Il mulino, Bologna 1969.
- A. Menichetti, *Metrica italiana: fondamenti metrici, prosodia, rima*, Antenore, Padova 1993.
- B. Mortara Garavelli, *Manuale di retorica*, Bompiani, Milano 1997.
- M. P. Ellero - M. Residori, *Breve manuale di retorica*, Sansoni, Milano 2001.
- S. Ghiazza - M. Napoli, *Le figure retoriche: parola e immagine*, Zanichelli, Bologna 2007.
- B. Mortara Garavelli, *Il parlar figurato. Manuale di figure retoriche*, Laterza, Bari 2010.
- S. Arduini - M. Damiani, *Dizionario di retorica*, LacCom 2010.
- P. G. Beltrami, *La metrica italiana*, Il mulino, Bologna 2011.
- G. Lavezzi, *Breve dizionario di retorica e stilistica*, Carocci, Roma 2017.
- U. Motta, *Lingua mortal non dice: guida alla lettura del testo poetico*, Carocci, Roma 2020.
- TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange* [4.4.0] [19.04.2022].

Closing Note

Thank you for coming to TEI2022 (if you did), and by the end of it we hope everyone had an enjoyable, fruitful, and engaging experience. It has been our pleasure to host you in Newcastle Upon Tyne at Newcastle University.

This document, along with a variety of other materials will be deposited at:

<https://zenodo.org/communities/tei2022>

for long-term preservation. We hope that this will be useful to those interested in the #TEI2022 conference and the Text Encoding Initiative more generally.

Dr James Cummings
Lead Local Organiser
Newcastle University
September 2022