



**Text Encoding Initiative Conference
and Members' Meeting 2025**

Kraków, Poland
September 16-20, 2025

Book of Abstracts

Wydział Filologiczny, Uniwersytet Jagielloński
al. Adama Mickiewicza 9, Kraków

**“New Territories”. Text Encoding
Initiative Conference and Members’
Meeting 2025. September 16–20,
2025. Kraków, Poland.**

Book of Abstracts

Edited by:

Joanna Halaczekiewicz

Organized by:

Jagiellonian Centre for Digital Humanities

al. Mickiewicza 11, 31-120 Kraków

jchc@uj.edu.pl

tei2025.confer.uj.edu.pl

ISBN: 978-83-977695-0-2

DOI: 10.5281/zenodo.17312233

Layout and cover design: [Patrycja Wojkowska](#)

Typesetting: [Joanna Halaczekiewicz](#)

Proofreading: Kinga Moskał, Wiktoria Szelest

Disclaimer

The content of each abstract is the sole responsibility of its author(s). The editors and organizers are not responsible for the views expressed in the abstracts.

License

This publication is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

Proofreading of this publication has been supported by a grant from the Priority Research Area (PRA Heritage) under the Strategic Programme Excellence Initiative at Jagiellonian University.

Table of Contents

7	Introduction
9	From the editors
11	Keynotes
13	A Case for Best Practices in TEI Encoding of Newspapers. The Darmstädter Tagblatt as an example
15	A new attribute class for annotating syntactic dependency relations
20	A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents
28	An XML-based edition publication model
31	Between Language and Music. A Case Study on Encoding of Textbooks on Japanese Traditional Music
33	Beyond Rule-Based Processing. LLM-Assisted TEI Encoding of Editorial Interventions in Historical Correspondence
37	Can we make an AI respect TEI XML? An Experiment with a Small-Scale Explainable AI Model
40	Converting data from Czech digital libraries to TEI
44	Differentiating ODDs
46	Dispatches from the TEI's GIS Working Group
48	Editing Multilingual Grand Vizierial Correspondence in TEI. The Graviz project
50	Encoding a Work of Interactive Fiction. TEI, Flash, and the Afterlives of ApertureScience.com
53	Encoding Ambiguity. A TEI Approach to Argentinean Reading Theory in an 80s Literary Magazine
55	Encoding Argentinian Eighteenth and Nineteenth-Century Drama with TEI
58	Encoding Complexity. TEI Modeling in the “Forschungsportal BACH” Project
60	Encoding Hyperfiction. Preliminary Considerations for a Digital Edition of Susanne Berkenheger’s “Hilfe! Ein Hypertext aus vier Kehlen” (1998)

- 63 Encoding language diversity in TEI. A description of regional and non-standard languages in multilingual diachronic corpora
- 68 Encoding Ligature Glyphs in Arabic Manuscripts Using TEI-XML
- 70 Encoding Multilingual and Multiscript Sources in TEI. A historical Spanish-Chinese dictionary
- 76 Enriching textual data from digital libraries
- 79 Exploring correspSearch v3 as a Service for Minimal Editions of Correspondence
- 84 Eye Rhyme in the Digital Victorian Periodical Poetry Project
- 87 FAIR research data from Goethe's inbox: The first 2,400 incoming letters as TEI-XML full texts
- 93 Foregrounding Text Analysis in Digital Scholarly Editing: Who Are We Encoding For?
- 98 From dialect features to structured data. Modelling spoken Arabic varieties in the WIBARAB project
- 102 From Practice to Framework. Model Digital Scholarly Editions in the Jagiellonian Digital Platform
- 106 Functors and format conversion
- 109 Generating TEI Documents Through a Game of Dice
- 111 Generative AI for XML-TEI Encoding
- 114 Graph-Based Digital Editions and TEI. Towards Interoperable and Assertive Scholarly Editing
- 119 Haute couture (for the masses)
- 121 Introducing UDraCor. The Ukrainian Drama Corpus in TEI
- 124 Ioannes Dantiscus' Itinerary Online
- 126 Jinks / TEI Publisher 10
- 128 JinnTap. A browser-based TEI-XML editor
- 130 LEGOstyle. Building modular, flexible Editions with TEI
- 134 Linking Your *-ographies. Developing project-specific TEI Authority File Lookups for LEAF-Writer
- 136 Managing lexical complexity. ODD chaining for Arabic dialect dictionaries
- 140 Multimodality and Minimal Publishing. TEI, MEI, and more in 19th-Century Music Treatises
- 145 Navigating and Processing Data from the TEI with XPath and XSLT
- 147 New features of Scholarly XML, an Open Source Visual Studio Code Extension for TEI encoding
- 150 New Territories for (Very) Old Language: DiaCorPolL. Automated Compilation of the Diachronic Corpus and Dictionary of Latin in Poland
- 154 New Territories for Digital Lexicography. Dictionary of Polish Dialects meets TEI
- 157 ODD-API. A REST Interface for Programmatic Use of ODD Schema Definitions
- 161 Parallel Editing in TEI. The Case of Regestra 1561
- 163 Processable and Computable Media in TEI. Usecases and Strategies
- 167 Reconstructing the Experience of a Historian from the Meiji (明治) Period. Attempts to Encode Kengo Murakawa (村川堅固)'s Diary in TEI
- 171 Save it on a Shoe String. Experiences with Project Endings
- 173 Standardisation and innovation. The role of TEI in improving innovative potential in digital scholarly editions of correspondences
- 175 Starting from Showing, Moving toward Sharing. Broadening TEI's Reach
- 177 Streamlining TEI Workflows. Collaborative Editing with NormaTEI
- 183 Structural Markup of the Database of Historical Polish Lexicons. The Case of *Forytarz języka polskiego* by Jan Ernesti and *Nowy dykjonarz, to jest Mownik polsko-niemiecko-francuski* by Michał Abraham Troc
- 186 Structures and Tools for the Representation and Visualization of Knowledge Contained in Electronic Editions/Textual Resources
- 188 TEI Processing Model
- 190 The GMIerator. A generalised pipeline for contributing to correspSearch

193	The TEI Critical Apparatus at 30-something. Where do we go from here?
195	Towards a model of transcultural encoding of ancient epigraphic sources
198	Towards an Encoding Practice for Multilingual Textual Variation
201	Voci dall'Inferno. A TEI-Based Digital Archive for finding Dante in Concentration Camp Testimonies
205	What do we Need to Make Documents from Texts?
209	Who Knows What a Revision Is? Towards a Shared Vocabulary of Textual Variation
211	Program Committee
212	Local Organizers

Introduction

It was my great pleasure to see the 25th Annual Meeting of the Text Encoding Initiative—my beloved scholarly community—take place in Kraków, my hometown and favorite city. On paper, TEI is a technical standard and a non-profit organization. In reality, “*we are the TEI*” (as the title of our closing session with Julia Flanders states). It’s a network connected not so much by formal membership but rather by curiosity and a willingness to explore the fertile crescent where diverse intellectual cultures meet.

This diversity is clearly reflected in this year’s program, with subjects ranging from frameworks and interfaces to linguistics; from music and literature to the conceptual and technical underpinnings of TEI itself; from nuanced representations of irony to accountable AI.

Is throwing such a wide spectrum of scholarly interests together a feature or a bug? Quoting the opening paragraph of the TEI website, “*we develop the Guidelines, which provide the infrastructure for developing machine-actionable cultural heritage texts.*” Therefore, this richness is exactly what TEI was designed for: encoding the world, not just the words; providing a platform for hybrid voices and approaches—historical sources and new interpretations, standardization and heterogeneity— all at the same time.

This is also what I return to every year at the TEI meeting: great, generous minds and ODDs (forgive the pun), always in favor of lively debate, finding intellectual inspiration, and sowing seeds for fruitful collaborations in the future.

This book of abstracts can only be an echo of the event; nevertheless, it gives a taste of some of the good things we enjoyed.

Magdalena Turska

Chair of the Programme Committee

From the editors

Dear Readers,

as you flip through this Book of Abstracts, you will encounter voices from across the TEI community exploring textual scholarship, markup, digital editions, and the infrastructures that support our shared work. The theme *New Territories* gestures toward both geographic and intellectual horizons, encouraging us to discover together how TEI may unfold in new places, contexts, practices, and collaborations.

Kraków, with its rich history and lively academic community, was a fitting place for our meeting. As Poland's oldest university, the Jagiellonian offers both a chance to reflect on tradition and an opportunity to engage with digital scholarship in fresh ways.

This volume is more than a program digest: it is a map of current experiments, emerging standards, and conversations that point to where TEI might go next. You will find abstracts addressing interoperability, machine learning for encoded texts, multimodal markup, regional efforts in underrepresented contexts, sustainable infrastructure, and more – reflections of the community's shared commitments.

We are enormously grateful to the authors, reviewers, organizers, and volunteers who shaped this program; and above all to you – the participants – for bringing your energy, insights, and openness to collaboration.

The conference was a true space of discovery, dialogue, and new connections – a moment when territory became not a limit but a horizon.

With best wishes,

Magdalena Komorowska

on behalf of the TEI 2025 Local Organizing Committee

Keynotes

Perspectives on New Territories in
Digital Editing

Jadwiga Kita-Huber

Jagiellonian University

An Introduction to the European
Research Council and why it
matters for the DH/TEI community

Sebastian Winkler

European Research Council

We are the TEI

Julia Flanders

Northeastern University

A Case for Best Practices in TEI Encoding of Newspapers

The Darmstädter Tagblatt
as an example

Kevin Kuok

ULB Darmstadt, Germany

Keywords

TEI encoding, newspapers and periodicals,
best practices, AI, digitization

Abstract

This poster addresses the reuse of periodicals and newspapers and proposes a best practice for TEI encoding in digitization projects, focusing on the Darmstädter Tagblatt. As part of the newspaper working group of DHd - Association for Digital Humanities in the German Speaking Areas, we recognize the growing need for standardized TEI encoding to facilitate data reuse across various newspaper projects. Leveraging insights from a recently initiated series of workshops on the reuse of newspaper data, hosted in Darmstadt in 2024 and in Vienna in May 2025, the objective is to explore the potential of TEI in enhancing access to historical newspaper data while fostering collaboration among researchers. As the project advances, a range of ideas is taking shape. One such idea is the creation of a universal TEI header that is suitable for both newspaper and periodical editions. An ongoing digitization project at

ULB Darmstadt, which is funded by DFG (German Research Foundation), will serve as the scientific basis. The poster will provide an overview of the project “a Darmstadt Newspaper in Three Centuries - Digitization of the Darmstädter Tagblatt, 1740 – 1986, (Thomas Stäcker, Marcus Müller, Dario Kampkaspar, et. al.)”, highlighting its significance, which also extends to AI/ML research. At TU Darmstadt, an AI assistant is being developed using one of Germany’s longest-running periodicals as a basis. It represents a heterogeneous data set and provides an excellent case study for establishing a best practice. It will also address the challenges and opportunities in TEI encoding of newspapers, proposing general best practices. Engaging with the international TEI community, we seek to foster discussions on standardization, collaborative markup, and the revitalization of the SIG “Newspapers and Periodicals.” By establishing standardized TEI encoding practices for newspapers, we aim to facilitate collaboration, enhance access to resources, and advance research in digital humanities.

A new attribute class for annotating syntactic dependency relations

Piotr Banski (1);
Andreas Nolda (2);
Harald Lüngen(1)

(1) IDS Mannheim, Germany;
(2) Berlin-Brandenburg Academy
of Sciences and Humanities
(BBAW), Germany

Keywords

dependency annotation, lightweight
grammatical annotation, Universal
Dependencies

Abstract

Overview

Language corpora, used as data for linguistic research and machine learning, are traditionally annotated with lemmas, part-of-speech tags, and, possibly, morphosyntactic categories. As a lightweight, inline, representation of such (mostly automatic) annotations, the TEI Guidelines provide the att.linguistic class with `@lemma`, `@lemmaRef`, `@pos`, `@msd`, and `@join` attributes (cf. Bański et al., 2018). Increasingly, corpora are also annotated with syntactic parses in terms of dependency relations between tokens, as a basis for syntactic queries or inferences. A de facto standard for that, supported by a wide array of tools, is the Universal Dependencies framework (UD; cf. Marneffe et al., 2021) with its CoNLL-U format. The present contribution outlines a potential extension of the TEI Guidelines for annotations of syntactic dependency relations, in the form of a new attribute class

called `att.linguistic.dependency`, which extends `att.linguistic` with the attributes `@head` and `@deprel` and several conventions.

Example

As a rule, syntactic dependency relations between tokens can be modelled by annotating each token with a reference to its syntactic head (if any) and with a label that states the linguistic type of the dependency. Such an approach also underlies the plain-text, column-based CoNLL-U format, where each token is annotated, *inter alia*, with a numerical index in the ID column, a reference to the index of its head in the HEAD column, and a label of the type of the dependency relation in the DEPREL column. The proposed addition makes it possible to represent this kind of information in TEI XML, as in the following example.

```
<s>
  <w n="1" head="2" deprel="nsubj" pos="PRON"
lemma="she">She</w>
  <w n="2" head="0" deprel="root" pos="VERB"
lemma="buy">buys</w>
  <w n="3" head="2" deprel="obj" pos="NOUN"
lemma="book"

  join="right">books</w>
  <pc n="4" head="2" deprel="punct" pos="PUNCT"
lemma=".">.</pc>
</s>
```

Proposed TEI encoding of syntactic dependency relations in the sentence *She buys books*. Note that the `<s>` element is not the only container possible.

The tokens are marked up with `<w>` tags for words and `<pc>` tags for punctuation characters, the global `@n` attribute is used for the numerical token index, the proposed `@head` attribute holds the numerical index of its head, and the proposed `@deprel` attribute labels the type of the dependency relation. The root node

(typically, but not necessarily, a verb) may either be marked up implicitly as in traditional dependency syntax (cf. the overview in Heringer, 1993) by omitting `@head` and `@deprel` attributes, or explicitly as in CoNLL-U with a `@head` value "0" and a `@deprel` value "root". In typical use cases, also lemmas, parts-of-speech tags, and, possibly, morphosyntactic categories are annotated using the `@lemma`, `@pos`, and `@msd` attributes, corresponding to the CoNLL-U columns LEMMA, UPOS, and FEATS, respectively; the `@join` attribute signals that a token is directly adjacent to another.

The proposed lightweight annotation scheme is also capable of representing more complex syntactic annotations, involving concurrent part-of-speech tagging (CoNLL-U column XPOS) or enhanced syntactic dependencies (CoNLL-U column DEPS) as they are used in UD for coordination ellipsis and related syntactic phenomena. The 'vanilla' as well as advanced uses of the proposed class will be discussed and illustrated during the presentation.

Summary

The proposed attribute class `att.linguistic.dependency` with the attributes `@head` and `@deprel` extends the lightweight means provided by the TEI Guidelines for encoding linguistic corpus annotations in TEI XML from the level of token-related properties (defined mostly by `att.linguistic`) to the level of inter-token relationships.

It may be used to extend existing collections of TEI-encoded texts by syntactic dependency annotations, as well as in dedicated NLP pipelines for new annotated corpora. Being lightweight by design, it does not pretend to be a replacement for full-fledged structural and grammatical descriptions enabled by specialised TEI components, such as those underlying standoff representations (Guidelines, ch. 17.10), graph representations (Guidelines, ch. 20.1), tree-based constituent structure representations (Guidelines, ch. 20.3 and the `att.segLike` class with

<s>, <cl>, <phr>], or robust descriptions of grammatical structures and content by means of the mechanisms of feature structures (FSR, cf. Guidelines, ch. 19). Rather, it constitutes a step towards opening the TEI to those projects that need to create syntactically annotated resources quickly, using the currently available mainstream tools, but do not wish to be confined to a plain-text format.

References

Draft spec for att.linguistic.dependency: <https://jenkins-paderborn.tei-c.org/view/LingSIG/job/TEIP5-LingSIG-att.linguistic.dependency/lastSuccessfulBuild/artifact/P5/Guidelines-web/en/html/ref-att.linguistic.dependency.html>

Draft addition to Chapter 18.3, “Syntactic dependency relations between word-level elements”: <https://jenkins-paderborn.tei-c.org/view/LingSIG/job/TEIP5-LingSIG-att.linguistic.dependency/lastSuccessfulBuild/artifact/P5/Guidelines-web/en/html/AI.html#AILADEP>

Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight Grammatical Annotation in the TEI: New Perspectives. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1283/>

Hans Jürgen Heringer. 1993. Basic Ideas and the Classical Model. In: Syntax: Ein internationales Handbuch zeitgenössischer Forschung/An International Handbook of Contemporary Research, ed. by Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Vennemann, Handbücher zur Sprach- und Kommunikationswissenschaft 9/1, Berlin: de Gruyter, 298–316. Available at <https://www.degruyterbrill.com/serial/hksyn-b/html?lang=en>

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics 47(2): 255–308. https://doi.org/10.1162/coli_a_00402

A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents

Juliette Janès,
Sarah Bénére,
Benoît Sagot,
Thibault Clérice

Inria, Paris, France

Keywords

Automatic encoding, Layout analysis,
Annotation guidelines

Abstract

Thanks to Automatic Text Recognition (ATR) workflows, we are able to automatically process images of textual documents into TEI-encoded files (Pinche and Spychala, 2024). Those workflows often rely on existing controlled vocabularies like SegmOnto (Gabay, Camps, and Pinche 2021) for layout analysis tasks such as identifying the structural components of a page (e.g. titles, page numbers, columns). However, the resulting TEI output offers only a shallow representation of the document's structure. It is mostly limited to a basic separation between the main body of text and elements considered as "noise" (e.g. running titles, tables, etc.). In order to address this gap, we present

the LADaS¹ Annotation Guidelines—building on both SegmOnto and the TEI—and their associated dataset (Appendix 1). By using a model trained on LADaS, we aim at automatically reconstructing documents in greater depth by adding more detailed description levels aligned with the TEI Guidelines.

We use a system of two-level annotations (Appendix 2):

1. A semantic level describing zones that convey a specific meaning on the page, like running titles, numbering, marginal notes, or the main body of text;
2. A graphical sublevel describing the visual aspect of a sub-zone, like paragraphs, list items, or headings. Second-level elements are never found on their own and can be repeated throughout the zones to specify their content (e.g. margin text notes can be composed of multiple paragraphs).

The semantic level is based on SegmOnto and describes the different zones of the layout. While SegmOnto is modelled from the content of manuscripts and early prints,² we added new zones like FigureZone (for code snippets), to deal with more recent documents. The semantic level encompasses media (GraphicZone, TableZone, FigureZone and FormZone), and makes a distinction between primary text zones (MainZone, MarginTextZone and TitlePageZone)—which are generally specified by an element from the second level—and liminal text zones (DigitizationArtefactZone, NumberingZone, RunningTitleZone, StampZone and QuireMarksZone)—which are never specified further.

The graphical sublevel focuses on the visual description of the subzones like indentation, line types, typographic features, and text structuring. The sublevel is aligned as much as possible

¹ For more information on LADaS, see <https://github.com/DEFI-COLaF/LADaS/tree/main>.

² Based on the authoritative works developing specialised vocabulary like Codologia (<https://codicologia.irht.cnrs.fr/>).

with existing TEI elements. As a result, `<head>`, `<p>`, `<quote>`, `<lg>`, `<signed>`, `<dateline>`, and `<address>` are equivalents in both the TEI and LADaS, but others are too specific and cannot be defined by TEI definitions as they are. For instance, we have identified five graphic variations for paragraphs each having its associated TEI post-processing element.³

Therefore, while most of the TEI file can be automatically generated using the zoning information provided by the annotations, additional post-processing is sometimes still needed. We introduced Continued—a non-TEI element—to handle interrupted textual content. This specific annotation is used when the text appears at the top of the MainZone or is interrupted by another zone (e.g. an image or table), making it impossible to determine the exact second-level element it belongs to. During post-processing, we attempt—whenever possible—to associate this portion of text with its preceding zone, so as to reconstruct the document in TEI as accurately as possible.

³ P (basic prose paragraph, `<p>`), P-Labelled (paragraph with a form of heading, `<p rend="labelled">`), P-Structured (entries, `<p rend="structured">`), P-Styled (entirely bold or italic paragraph, `<p rend="styled/bold/italic">`), P-Quoted (indented quotations, `<quote>`).

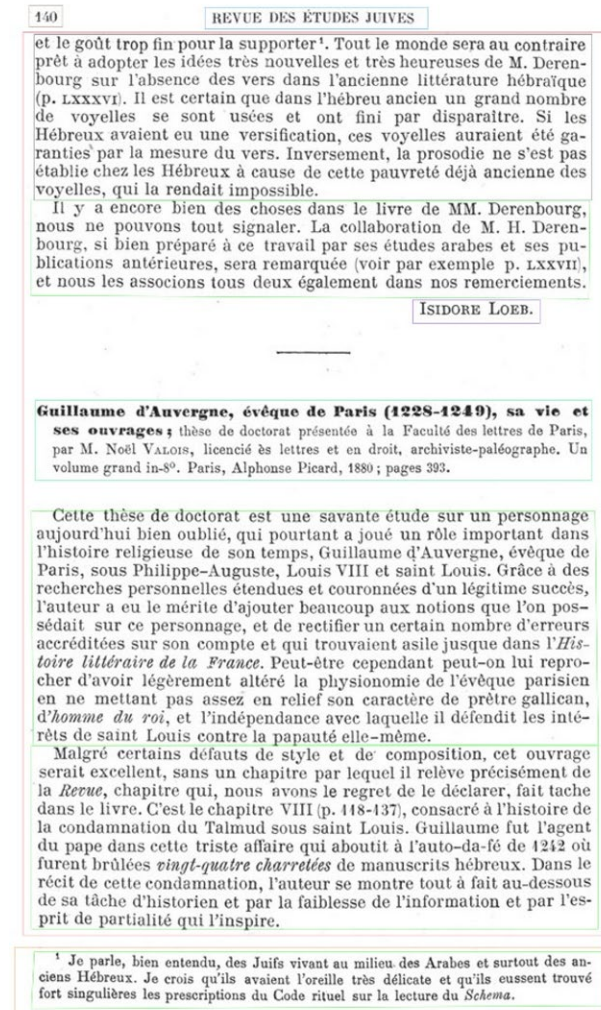


Fig. 1: Example of an annotated image.

(1st Level Annotations: MainZone in red, MarginTextZone in orange, NumberingZone in grey and RunningTitleZone in blue; 2nd Level Annotations: -P in green, -P-Structured in light green, -Signed in violet and -Continued in black)

```

<div>
  <!--...-->
  <pb n="140"/>
  <fw type="numbering"><lb/> 140</fw>
  <fw type="runningtitle"><lb/> REVUE DES ÉTUDES JUIVES</fw>
  <ab rend="continued"><lb/> et le goût trop fin pour la supporter!. Tout le monde sera
  au contraire <lb/> prêt à adopter les idées très nouvelles et très heureuses de
  M. Deren- <lb/> bourg sur l'absence des vers dans l'ancienne littérature
  hébraïque, <lb/> (p. LXXXVI). Il est certain que dans l'hébreu ancien un grand
  nombre <lb/> de voyelles se sont usées et ont fini par disparaître. Si les <lb/>
  Hébreux avaient eu une versification, ces voyelles auraient été gar- <lb/> ranties
  par la mesure du vers. Inversement, la prosodie ne s'est pas <lb/> établie chez les
  Hébreux à cause de cette pauvreté déjà ancienne des <lb/> voyelles, qui la
  rendait impossible.</ab>
  <p><lb/> Il y a encore bien des choses dans le livre de MM. Derenbourg, <lb/> nous ne
  pouvons tout signaler. La collaboration de M. H. Deren- <lb/> bourg, si bien
  préparé à ce travail par ses études arabes et ses pu- <lb/> blications
  antérieures, sera remarquée (voir par exemple p. LXXVII), <lb/> et nous les
  associons tous deux également dans nos remerciements.</p>
  <signed><lb/> ISIDORE LOEB.</signed>
</div>
<div>
  <p rend="structured">
    <lb/> Guillaume d'Auvergne, évêque de Paris (1228-1219), sa vie et <lb/> ses
    ouvrages; thèse de doctorat présentée à la Faculté des lettres de Paris, <lb/>
    par M. Noël Valois, licencié ès lettres et en droit, archiviste-paléographe. Un
    <lb/> volume grand in-8°. Paris, Alphonse Picard, 1880; pages 393.</p>
    <p><lb/> Cette thèse de doctorat est une savante étude sur un personnage <lb/>
    aujourd'hui bien oublié, qui pourtant a joué un rôle important dans <lb/>
    l'histoire religieuse de son temps, Guillaume d'Auvergne, évêque de <lb/> Paris,
    sous Philippe-Auguste, Louis VIII et saint Louis. Grâce à des <lb/> recherches
    personnelles étendues et couronnées d'un légitime succès, <lb/> l'auteur a eu le
    mérite d'ajouter beaucoup aux notions que l'on pos- <lb/> sédait sur ce personnage,
    et de rectifier un certain nombre d'erreurs <lb/> accréditées sur son compte et qui
    trouvaient asile jusque dans l'His- <lb/> toire littéraire de la France. Peut-être
    cependant peut-on lui repro- <lb/> cher d'avoir légèrement altéré la physionomie
    de l'évêque parisien <lb/> en ne mettant pas assez en relief son caractère de
    prêtre gallican, <lb/> d'homme du roi, et l'indépendance avec laquelle il défendit
    les inté- <lb/> rêts de saint Louis contre la papauté elle-même.</p>
    <p><lb/> Malgré certains défauts de style et de composition, cet ouvrage <lb/> serait
    excellent, sans un chapitre par lequel il relève précisément de <lb/> la Rerue,
    chapitre qui, nous avons le regret de le déclarer, fait tache <lb/> dans le livre.
    C'est le chapitre VIII (p. 118-137), consacré à l'histoire de <lb/> la condamnation
    du Talmud sous saint Louis. Guillaume fut l'agent <lb/> du pape dans cette triste
    affaire qui aboutit à l'auto-da-fé de 1242 où <lb/> furent brûlées vingt-quatre
    charretées de manuscrits hébreux. Dans le <lb/> récit de cette condamnation,
    l'auteur se montre tout à fait au-dessous <lb/> de sa tâche d'historien et par la
    faiblesse de l'information et par l'es- <lb/> prit de partialité qui l'inspire.</p>
    <note><p><lb/> 1 Je parle, bien entendu, des Juifs vivant au milieu des Arabes et
    surtout des an- <lb/> ciens Hébreux. Je crois qu'ils avaient l'oreille très
    délicate et qu'ils eussent trouvé <lb/> fort singulières les prescriptions du
    Code rituel sur la lecture du Schema.</p></note>
  </p>
</div>

```

Fig. 2. Suggested TEI encoding produced using the annotations

References

- Clérice, Thibault, Juliette Janès, Hugo Scheithauer, Sarah Bènière, Laurent Romary, and Benoît Sagot. 2024. "Layout Analysis Dataset with SegmOnto." Presented at DH 2024 - Reinvention and Responsibility, Washington, D.C., 7-9 August 2024. URL: <https://inria.hal.science/hal-04513725>.
- Gabay, Simon, Jean-Baptiste Camps, and Ariane Pinche. 2021. "SegmOnto: Un vocabulaire contrôlé pour décrire la page manuscrite et imprimée." Presented at Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le Xe-XIVe siècle, Paris, 9 November 2021. URL: <https://hal.science/hal-03481089v1/file/SegmOnto.pdf>.
- Pinche, Ariane, and Pauline Spychala. 2024. "Getting Started with Automatic Text Recognition," Automatic Text Recognition: Harmonising ATR Workflows (blog), 7 May 2024. DOI: 10.58079/11npw.
- TEI Consortium (Eds). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.9.0. Last modified 24th January 2025. TEI Consortium. URL: <https://tei.org/release/doc/tei-p5-doc/en/html/index.html>.

Appendix 1. Overview of the LADaS Subsets

Subset	Century	Number of Images
Administrative Reports	19-21	229
Catalogues	19-20	1,437
Fingers	21	100
Magazines	20-21	330
Monographies	17-20	1,992
Others	21	6
Persée	20	1,230
Picard	21	97
Romans 19	19	240
Theatre	17-20	750
Theses	21	745
Typewriter	20	98
TOTAL		7,254

Appendix 2. LADaS Annotation Classes

Semantic level (SegmOnto) Graphical level (TEI)

MainZone	-Head
MarginTextZone	-P
TitlePageZone	-P-Labelled
GraphicZone	-P-Structured
FigureZone	-P-Quoted
TableZone	-P-Styled
FormZone	-Item
DigitizationArtefactZone	-Lg
NumberingZone	-Dateline
RunningTitleZone	-Address
StampZone	-Signed
QuireMarksZone	-Ab
	-Continued
	-Part
	-Decoration
	-Maths
	-Field
	-DropCapital

An XML-based edition publication model

Peter Boot

The Huygens
Institute, Netherlands

Keywords

publication, XML, specification

Abstract

This paper proposes an XML-based Edition Publication Model (EPM). While there now exists a number of mature tools for publishing TEI-encoded texts, such as TEI Publisher, Edition Visualisation Technology and CETEIcean, the paper will argue there is room for a software-independent specification of the functionality of digital editions. Just as the TEI itself was among other things an attempt to move away from software-specific encoding of textual features (Ide & Sperberg-McQueen 1995), an Edition Publication Model should be able to describe how the edition data (TEI XML files) should be displayed and what interactional features this display should offer, without assuming a certain software context.

The paper will discuss (1) the reasons for this proposal, (2) the relation to similar initiatives and existing tools, (3) some aims that an EPM should fulfil and (4) a first schema and implementation of an EPM. The aim of the presentation is not just to inform the public, but also to attract potential collaborators.

Ad 1. Briefly: all software is temporary and eventually dies. Data that conforms to a public specification is likely to outlive

the software that handles it. If an edition's interface dies but the definition of the interface is still available, it is easier to recreate that interface in another tool. Besides, if there is a public specification of edition functionality, it becomes easier to write software that can display multiple editions out of the box, rather than the custom-written software that we often see today.

Ad 2. While there is a longer history of related initiatives, the ones that the paper will highlight are the Ediarum manifest (Fechner 2018) and the TEI Processing Model (Turska, Cummings & Rahtz 2016). Ediarum manifests provide a high-level definition of some aspects of the edition but leave other aspects to the implementation and the details of processing to XSLT stylesheets. The TEI Processing Model describes a set of edition behaviours and provides tools to map XML structures to these behaviours. The paper will argue an extension of the TEI Processing Model is an essential ingredient for an EPM (Boot 2024).

Ad 3. The EPM should facilitate a complete specification of the content and behaviour of an edition. As new edition functionalities continue to appear (think of Pure3D⁴ or editions that integrate network displays⁵), EPM should be extensible. It should also be modular: it should be possible to model the several components that an edition can contain (say charters, personographies, introductions) as well as the interaction between these components. Finally it should be chainable: it should be possible to override behaviour mappings defined at a general level with other mappings useful in specific situations.

Ad 4. Finally, the paper will describe a first iteration of a schema and an XSLT-based implementation for an EPM.⁶ This first iteration only aims to describe data display, and is not concerned

4 See e.g. <https://editions.pure3d.eu/project/14/edition/1/>.

5 See e.g. <https://db.innovatingknowledge.nl/edition/>.

6 See <https://gitlab.huc.knaw.nl/edition-publication-model/edition-publication-model>.

with interaction or searching. The discussion will present a number of design issues for an EPM.

References

- Boot, P. (2024). The TEI Processing Model: Introduction, limitations and potential extensions. In Declarative Amsterdam 2024. <https://doi.org/10.1075/da.2024.boot.tei-processing-model>.
- Fechner, M. A (2018). Standardized Interface for Digital Scholarly Editions. DHd 2018. <https://edoc.bbaw.de/frontdoor/index/index/docId/3327>.
- Ide, N. M., & Sperberg-McQueen, C. M. (1995). The TEI: History, goals, and future. *Computers and the Humanities*, 29, 5-15.
- Turska, M., Cummings, J., & Rahtz, S. (2016). Challenging the myth of presentation in digital editions. *Journal of the Text Encoding Initiative*, (9).

POSTER

Between Language and Music

A Case Study on Encoding of Textbooks on Japanese Traditional Music

Shintaro Seki

RIKEN, Japan

Keywords

traditional Japanese music, musical notation, text and character

Abstract

The Text Encoding Initiative (TEI) has developed and maintained guidelines for the digital encoding of literary and linguistic texts. However, some materials use characters not as symbols for constructing literary or linguistic works, but for other purposes. This poster examines musical scores used in traditional Japanese music as a representative example of such materials and explores possibilities for their digital encoding.

Musical scores are documents designed to record music and convey essential information for performance through various types of symbols. In Western musical traditions, a specialized notation system has been developed to visually represent musical structures. In contrast, in Japanese musical culture, characters used in everyday language are often repurposed for musical notation. These are frequently syllables—known as *kuchishōga* in Japanese—sung to aid in memorizing melodies but

generally lacking inherent linguistic meaning. Furthermore, musical scores may also include lyrics and explanations of instrumental techniques. As a result, musical symbols and linguistically meaningful text coexist within the same character system.

This poster presents a case study of a music textbook that incorporates both musical symbols and linguistic text, focusing on how its content can be encoded using XML. It explores encoding strategies that accommodate the simultaneous use of symbols from the same character system for multiple functions within a single document, with particular attention to the coexistence of musical notation, semantic content, and accompanying text such as lyrics and performance notes. Additionally, the poster discusses current challenges in linking other notation systems, such as five-line staff notation transcribed using the Music Encoding Initiative (MEI) or Music XML, and linking text materials with time-series media such as audio and video recordings.

Beyond Rule-Based Processing

LLM-Assisted TEI Encoding of Editorial Interventions in Historical Correspondence

Sabrina Strutz, Martina Scholger,
Georg Vogeler

University of Graz,
Austria

Keywords

Large Language Models, TEI Encoding,
Digital Correspondence Editions,
Automated Annotation, Editorial
Interventions

Abstract

Editions originating from print or created within the print tradition require specific post-processing to fully realize their potential in the digital realm. This requires critical reinterpretation of editorial conventions which are often implicitly used by the editors and not fully explained in the editorial. The recording of editorial interventions is a central task in the digital edition pipeline for the creation of a critical text (Pollin et al, 2025). While XML-TEI offers a flexible framework for encoding such editorial interventions (TEI Consortium, 2025), actual practice—particularly in printed editions—often suffers from inconsistency and ambiguity. This paper examines the potential of transformer-based language models to resolve editorial ambiguities through contextual analysis. One illustrative case is the use of square brackets

representing different editorial actions as in H[ochwohlgeboren] (abbreviation), emporschwin[gen]des (addition), welches[sic] (indicating errors in the original), [Lücke] (omission) or [?] (illegible content). This kind of ambiguous notations in print editions—including various kinds of brackets or verbal descriptions of textual phenomena in footnotes—pose even more challenges (Beyer 2018, 32; Heckmann, 2013; Leeb, 2013; Schwäbische Forschungsgesellschaft, 2001). The different functions of the editorial notation might be clear to a human reader without explanation, but challenge rule-based processing.

LLMs' ability to transform unstructured text into semi-structured formats and their potential to streamline complex tasks make them promising for Digital Humanities practices. Beyond encouraging results in annotation tasks using generative AI (Ding et al., 2023; Gilardi, Alizadeh and Kubli, 2023; Kuzman, Mozetič and Ljubešić, 2023), recent experiments suggest these technologies could effectively support the labor-intensive annotation phase in digital edition projects (Czmiel et al., 2024; Pollin, 2024; Scholger, Strutz and Pollin 2024).

We present a case study using the correspondence of Austrian orientalist and diplomat Joseph von Hammer-Purgstall (1774-1856), where over 3,500 letters from a print edition are being transformed into a digital edition reflecting current DH standards. Unfortunately, clearly defined editorial guidelines are absent or only partially recoverable. While structural elements of the letters can be quite reliably extracted from Word templates via TEIGarage and pattern matching in XSLT into TEI, the ambiguous and inconsistent use of square brackets poses a significant challenge requiring semantic understanding beyond deterministic methods.

Our approach extracts segments in square brackets including surrounding context, applies tailored prompts (zero-shot, few-shot) to proprietary and open-source LLMs (GPT-4o, Claude 3.7 Sonnet, OLMo 2), and converts results into TEI encoding (e.g., `<choice><abbr>H</abbr><exp>H<ex>ochwohlgebor-`

`en</ex></exp></choice>`). A comparative evaluation is based on a manually annotated gold subset.

Beyond technical implementation, we address crucial methodological questions: How can we prevent transformer-based models from “over-modernizing” (Scholger et al, 2025) historical content? How can we ensure encoding consistency across large datasets? This research contributes to emerging TEI practices by exploring how generative AI can complement human editorial work while maintaining scholarly standards—recognizing that as LLMs integrate into digital scholarship, their capabilities must be assessed not only for what they can resolve but also for what they might obscure.

References

- Beyer, B. (2018) Praktische Tipps für die Edition landesgeschichtlicher Quellen. Materialien der Historischen Kommission für Westfalen, Bd. 15. [https://www.lwl.org/hiko-download/HiKo-Materialien_O15_\(2018-03\).pdf](https://www.lwl.org/hiko-download/HiKo-Materialien_O15_(2018-03).pdf).
- Czmiel, A., Dumont, S., Fischer, F., Pollin, C., Sahle, P., Schaßan, T., Scholger, M., Vogeler, G., Roeder, T., Fritze, C., and Henny-Krahmer, U. (2024, February 21) ‘Generative KI, LLMs und GPT bei digitalen Editionen’. DHd 2024 Quo Vadis DH (DHd2024), Passau, Deutschland. <https://doi.org/10.5281/zenodo.10698210>.
- Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., and Bing, L. (2023, June) ‘Is GPT-3 a good data annotator?’ arXiv. <https://doi.org/10.48550/arXiv.2212.10450>.
- Gilardi, F., Alizadeh, M., and Kubli, M. (2023, July) ‘ChatGPT outperforms crowd workers for text-annotation tasks’, Proceedings of the National Academy of Sciences, 120. <https://doi.org/10.1073/pnas.2305016120>.

- Heckmann, D. (2013) 'Leitfaden Zur Edition Deutschsprachiger Quellen (13.-16. Jahrhundert)'. Preußenland 3:7-13.
- Höflehner, W., Wagner, A., Koitz-Arko, G., and Kowatsch, S. (2021) Joseph von Hammer-Purgstall 1774-1856 : ein altösterreichisches Gelehrtenleben : Eine Annäherung. (Vol. 93, Forschungen zur geschichtlichen Landeskunde Steiermark). Graz: ADEVA.
- Kuzman, T., Mozetič, I., and Ljubešić, N. (2023, March) 'ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification'. Tech. rep., arXiv. <https://doi.org/10.48550/arXiv.2303.03953>.
- Leeb, J. (ed.). (2013) Der Reichstag zu Regensburg 1556/57. 2 Bde. München: Oldenbourg Verlag.
- McGillivray, B., Poibeau, T., and Fabo, P. R. (2020) Digital humanities and natural language processing: "Je t'aime... moi non plus." Digital Humanities Quarterly, 14(2). <https://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html>.
- Pollin, C. (2024) 'Workshopreihe "Angewandte Generative KI in den (digitalen) Geisteswissenschaften" (v1.1.0)'. <https://doi.org/10.5281/zenodo.10647754>.
- Pollin, C., Fischer, F., Sahle, P., Scholger, M., and Vogeler, G. (2025, forthcoming) 'When it was 2024 - Generative AI in the field of digital scholarly editions', Zeitschrift für digitale Geisteswissenschaften.
- Scholger, M., Strutz, S., and Pollin, C. (2024) 'Empowering Text Encoding with Large Language Models: Benefits and Challenges'. Zenodo. <https://doi.org/10.5281/zenodo.13969082>.
- Scholger, M., Cugliana, E., Fischer, F., Pollin, C., Sahle, P., and Vogeler, G. (2025, April 28) 'Bias und Evaluation - Beiträge zur Kritik der digitalen Edition mit generativer KI'. Zenodo, <https://doi.org/10.5281/zenodo.15299432>.
- TEI Consortium, eds. (2025) '4.3.2 Simple Editorial Changes'. TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.9.0. TEI Consortium. <https://tei-c.org/Vault/P5/4.9.0/doc/tei-p5-doc/en/html/CO.html#COED>.

Can we make an AI respect TEI XML?

An Experiment with a Small-Scale Explainable AI Model

Elisa Beshero-Bondar, Hadleigh Bills,
Alexander Fisher

Penn State Erie,
USA

Keywords

AI, LLM, TEI XML, XPath, RAG

Abstract

My students and I are developing a project to see whether we can refine and customize a "DigitAI" (named for our "Digit" program at Penn State Behrend). We want our DigitAI to serve as a guide to TEI as well as the XML stack technologies used in our program and projects. We want our AI to be more reliable than previous experiences with LLMs in coding assistance, because (if we are successful) it will deploy XPath and XML-processing technologies. The challenge of our project is to introduce the meaningful hierarchy of XML and TEI to an AI language model trained on sequenced tokens and word vector embeddings. If successful, the DigitAI will be minimal in size and adaptable by others, but most importantly, we hope to learn from the experiment. We are designing an "explainable AI" system that gives humanities students and scholars agency in the training, evaluation, and application of the AI system, heeding the call for

explainable modeling work in the January 2023 special issue of IJDH (Ries).

In attempting this, we are learning from two previous efforts shared at the TEI 2024 conference. Martina Scholger and Mohamed Khemakhem each presented papers about creating an AI assistant for encoding documents in TEI, and each made use of Retrieval Augmented Generation or RAG to construct a local knowledge base for accessible to a large language model. Scholger spoke of delivering specialized resources from a scholarly edition project to enhance the knowledge base as well as prompt engineering strategies to improve results. Khemakhem supplied the TEI Guidelines as a document to augment a language model's training, but found significant problems in the model's responses due to the manner in which the Guidelines are delivered to the model via "chunking" their text—sometimes in the middle of a section or example. To address the problem that Khemakhem discussed at TEI 2024, we are investigating the parsing of XML as a resource for RAG. Thanks to Jang and Lee's successful efforts in an architectural studies context, where generative AI models were able to work with XML input of BIM (building information modeling) data, we are optimistic that we can introduce the XML of the TEI Guidelines to a generative AI model, or at least to control the chunking of the input data by following element structures.

We have chosen LLaMa, introduced by Meta in 2023, as a customizable small language model because its models are cross-platform compatible (Macs, PCs, Linux), provided that machines have sufficient GPU and 16 GB of RAM. We are experimenting with putting lxml etree and possibly SaxonC libraries into the Python script we are using to customize our model. We may be deploying XSLT in our pipeline to output digestible and coherent units of information to be introduced as a RAG resource. By the time of the TEI conference, I hope to report on our progress, share how far we have succeeded and/or failed in our efforts, and what we have learned so far.

References

- Hassan El-Hajj, Oliver Eberle, et. al. "Explainability and transparency in the realm of digital humanities: toward a historian XAI." *International Journal of Digital Humanities (IJDH)* (2023) 5:299-331. <https://link.springer.com/article/10.1007/s42803-023-00070-1> Accessed 2024-12-08.
- Khemakhem, Mohamed et. al. "Enhancing Technical Knowledge Acquisition with RAG Systems: The TEI Use Case." TEI 2024 Conference presentation at Buenos Aires. 2024. <https://zenodo.org/records/13988319> Accessed 2024-12-08.
- Jang, Suhjung and Ghang Lee. "Interactive Design by Integrating a Large Pre-Trained Language Model and Building Information Modeling" *Computer Science > Artificial Intelligence*. arXiv: <https://arxiv.org/abs/2306.14165>
- Llama Hub resources for building RAG applications. <https://llamahub.ai/>. Accessed 2024-12-08.
- Newtfire.org training resources. <https://newtfire.org/>. Accessed 2024-12-08.
- Ries, Thorsten, Karina van Dalen-Oskam, and Fabian Offert. "Reproducibility and explainability in digital humanities." *International Journal of Digital Humanities (IJDH)*. (2024) 6: 1-7. <https://doi.org/10.1007/s42803-023-00083-w> Accessed 2024-12-08.
- Scholger, Martina et. al. "Empowering Text Encoding with Large Language Models: Benefits and Challenges." TEI 2024 Conference presentation at Buenos Aires. 2024. <https://zenodo.org/records/13969082> Accessed 2024-12-08.
- TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange*. TEI P5 Version 4.8.1. Last modified 1 November 2024. <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

Converting data from Czech digital libraries to TEI

Boris Lehečka

Moravian Library in Brno,
Czech Republic

Keywords

digital library, natural language processing,
text enrichment, ALTO

Abstract

In the years 2019-2022, within the project „DL4DH – Development of tools for more effective use and mining of digital library data to strengthen digital humanities research“, a set of applications has been developed that uses (meta)data stored in the Kramerius system, designed for Czech digital libraries, and converts them (i.e. ALTO) into TEI format, while the text is enriched using external tools UDPIPE and NameTag. UDPIPE identifies sentences in the text and tokens within them, which it lemmatizes and annotates with morphological annotation, for 80 languages. NameTag tokenizes the text and recognizes named entities in it: persons, geographical places, institutions, time and numerical data, etc., for 20 languages.

The resulting TEI document is created by taking bibliography from MODS file and by merging three documents from partial transformations (from ALTO, NameTag and UDPIPE). The `<facsimile>` element identifies the recognized areas (`<zone>`) in the original image. The text is divided into paragraphs (`<p>`) and sentences (`<s>`), individual words (`<w>`) use attributes (`@lemma`,

`@msd`) for lemmas and morphological categories. The header of the final document contains the categories of the recognized entities (in the `<textClass>` element) as well as information about the tools used during the enrichment: name of the application, the language model used, and the licensing arrangements for the output (`<appInfo>` and `<availability>`).

The first problem in transforming and merging data arises from the fact that in the ALTO format, the input data is divided into blocks and lines but not into paragraphs, which makes it difficult to split the text into separate sentences for further analysis. ALTO may contain recognition errors (at the character level), which complicates lemmatization and morphological analysis. The first problem can be addressed by analyzing the input blocks (e.g., values for indentation of the first line or the whole block, first uppercase letter, hyphens at the end of lines, etc.) and grouping related lines into a paragraph.

Another problem is that the external applications UDPIPE and NameTag work not with XML on the input, but only with plain text (although it can have a structured form, e.g. vertical, CoNLL-U, etc.). The output is available as XML (for NameTag) or as structured plain text (NameTag, UDPIPE). In order to avoid problems with merging the output data because of different numbers of tokens, the UDPIPE service is first used to split the input text into tokens and to apply morphological analysis and lemmatization. The output in CoNLL-U format is later used as input for entity recognition.

When sending the data, it is also advisable to provide the external tools with information about the language of the analyzed text. This can be done using metadata about the publication or recognized language recorded in the ALTO format. Complications arise when there are isolated foreign expressions within a paragraph.

Within the *Libri augmentati* application (programmed in XProc 3.1 with individual researchers in mind), a set of XSLT and XQuery

transformations provides the conversion of input data (ALTO, NameTag XML, or CoNLL-U) into TEI format.

References

- Boris Lehečka, (2025). Libri augmentati [online]. Version 1.0.0. [computer software]. Brno: Moravian Library in Brno. [Viewed 11 July 2025]. Available from: <https://github.com/moravianlibrary/libri-augmentati>
- Lehečka, B., Novák, D., Kersch, F., Hladík, R., Bišková, J., Sekyrová, K., Válek, F., Vozár, Z., Bodnár, N., Sekan, P., Bežová, M., Žabička, P., Lhoták, M. and Straňák, P., (2022). *Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách* [online]. Knihovna AV ČR, v. v. i. [Viewed 11 July 2025]. Available from: <https://hdl.handle.net/11104/0335692>
- Library of Congress, (2023). ALTO: Analyzed Layout and Text Object [online]. *ALTO: Technical Metadata for Layout and Text Objects (Standards, Library of Congress)*. [Viewed 11 July 2025]. Available from: <https://www.loc.gov/standards/alto/>
- Straka, M., (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 197-207. [Viewed 11 July 2025]. Available from: <https://aclanthology.org/K18-2020/>
- Straková, J., Straka, M. and Hajič, J., (2019). Neural Architectures for Nested NER through Linearization. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 5326-5331. [Viewed 11 July 2025]. Available from: <https://aclanthology.org/P19-1527/>
- TEI Consortium, (2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange [online]. *The Text Encoding Initiative*

Consortium. [Viewed 11 July 2025]. Available from: <https://www.tei-c.org/Vault/P5/4.9.0/doc/tei-p5-doc/en/html/>

Walsh, N., Berndzen, A., Imsieke, G. and Siegel, E., (2025). XProc 3.1 [online]. *XProc - Specifications*. [Viewed 11 July 2025]. Available from: <https://xproc.org/specifications.html>

Differentiating ODDs

Syd Bauman (1);
Helena Bermúdez Sabel (2);
Martin Holmes (3)

(1) Northeastern
University, USA;
(2) JinnTec, Germany;
(3) University of Victoria,
Canada

Keywords | TEI, ODD, ATOP

Abstract

As we understand it, the original plan for the TEI/ODD system was for two types of user-facing ODD files:⁷ base ODDs which contain specifications for classes, elements, attributes, datatypes, and macros, likely arranged within modules; and customization ODDs which assert which specifications from the associated base ODD should be used (possibly with additional specifications) to create a particular markup language. These two types were to be differentiated by the presence of a `<schemaSpec>` element in the latter, and not in the former. But three realities impinge on that simple plan:

1. Many users of the TEI/ODD system (including two of the authors) have created stand-alone markup languages (i.e., languages written in ODD but not based on TEI) using a single ODD file that has a `<schemaSpec>` element. Such an ODD file is not a customization ODD, but nonetheless has a `<schemaSpec>`

⁷ The ATOP project has developed a terminology system to help clarify several different types of ODD file, as well as the processes that take place during ODD processing; this can be found on the project Wiki: <https://github.com/TEIC/atop/wiki/Terminology>. During the presentation, we will briefly explain this terminology.

2. Using the current Stylesheets the “compiled” ODD (the initial output of combining a customization ODD with a base ODD) contains a `<schemaSpec>` element. This compiled ODD file may be used to generate a customized schema (in Schematron and either RELAX NG, XSD, or DTD), but also may be used as the base for another customization, in which case it is not a customization ODD, but nonetheless has a `<schemaSpec>`.

3. ODD-chaining (the process of creating a sequence of successive customizations, where each stage is turned into the base ODD for the next stage) has become prevalent and useful to the community. This means that any individual customization may itself become a base ODD, while retaining its `<schemaSpec>`.

Thus we cannot use the mere presence of a `<schemaSpec>` to automatically differentiate between these two types of ODD files, which we need to be able to do to effect automatic ODD chaining. Our paper will discuss these issues, demonstrate a moderately complicated XSLT function that we *think* would do the job,⁸ and suggest a simple modification to the TEI scheme that would solve the problem: Making the `@source` attribute of `<schemaSpec>` a required attribute, and providing a special-purpose value (e.g. “tei:NONE”) for use when the ODD is a base ODD.

⁸ For the current version, see the `atop:is-base-odd()` function in XSLT/ modules/functions_module.xslt in the TEI/atop git repository.

Dispatches from the TEI's GIS Working Group

Joey Takeda (1);
Martin Holmes (2)

(1) Simon Fraser University,
Canada;
(2) University of Victoria, Canada

Keywords

GIS; geohumanities; interchange

Abstract

Created in November 2024, the GIS Working Group (GIS-WG) was charged by the TEI Technical Council with proposing modifications to the TEI schema to support more detailed, granular, and standards-based methods of encoding geospatial data in TEI, along with corresponding updates to the Guidelines prose. The full charge is available at <https://tei-c.org/activities/workgroups/gis-charge/>.

This presentation will share the current state of the GIS-WG's work, outlining the group's objectives, current recommendations, and the rationale behind them. Early in 2025, we devised a survey on current practices and preferences, which was completed by a number of community members particularly interested in this topic. We have also received feedback and discussion from the community during the March TEI Community Call.

Drawing on community feedback and informed by existing practices, the GIS-WG has now developed a proposal for schema modifications allowing support for GeoJSON, WKT, GML, KML, and other standards within TEI encoding. We have sought

to ensure that these proposed changes are sufficiently flexible to support widely-used standards, appropriately constrained to facilitate robust interchange, and backwards compatible with existing practices. We propose expanding `<geo>` to allow content in any of the supported geographic standards and creating a new element, `<geoDef>`, to allow for centralized definitions of geographic schemes. These two elements will both be members of a new class of attributes, which will specify the encoding scheme used, its version, and, to promote interchange, the level at which the geographic scheme is being implemented.

Editing Multilingual Grand Vizierial Correspondence in TEI

The GraViz project

Stephan Kurz;
Yasir Yilmaz;
Dimitra Grigoriou;
Nilab Saeedi;
Michael Vogelsberger

Austrian Academy
of Sciences, Austria

Keywords

Early modern history, TEI edition,
Multilingual, Non-Latin script, Ottoman
Empire, Habsburg Monarchy

Abstract

This poster charts new territories: Ottomanists and historians have studied source material available in Turkish and European archives for a long time. While research has produced numerous historiographical outcomes that draw on those sources, the number of scholarly editions of sources remained low as a consequence of the dispersal of sources and the specific challenges of multilingual textual transmission. The limitations to access such resources include that most existing collections provide little to no open source data.

The project ›The Ottoman Grand Vizierate (1560s to 1760s)‹ - GraViz in short - centers around the diplomatic correspondence of six selected grand viziers of the Ottoman Empire, preserved

in the archives in Istanbul and Vienna. The project's editorial mission is to apply state of the art TEI encoding to these letters, which survive in multiple languages: The GraViz team transcribes texts in Ottoman Turkish, Latin, and German language, and translates all Ottoman Turkish texts and a select group of German and Latin letters into English. Documents written in other languages, e.g. Italian and French, may follow in later stages, as well as detailed prosopographical information on the individual grand viziers.

The editorial team references named entities (persons, places) and dates (challenges in premodern calendar systems ensue), and provides keywords and an English-language abstract. CorrespDesc and other editorial metadata are included in the editions, along with facsimile images that show the challenging layout and reading order.

The project team published the first batch of edited documents, containing the correspondence of Sokollu Mehmed Pasha (in office: 1565-1579), at <https://qhod.net/context:graviz?locale=en>. The poster contribution focuses on how the GraViz project creates TEI XML data (schema, auxiliary standOff data, telHeader, alignment of original and translation/s), which serve as a solid digital foundation for future research on textual artifacts of diplomacy between the Ottoman Empire and European courts.

Encoding a Work of Interactive Fiction

TEI, Flash, and the Afterlives of ApertureScience.com

Alan Galey (1); Ellen Forget (1);
Brendan Allen (1);
Raffaele Viglianti (2)

(1) University of Toronto,
Canada;
(2) University of
Maryland, USA

Keywords

born-digital; interactive fiction;
transmedia; obsolescence

Abstract

How might the TEI Guidelines adapt to represent born-digital texts, such as works of interactive fiction? What parts of TEI P5 are already suited to this challenge? What does it mean to apply digital markup to texts that are already digital? And what productive debates about digital textuality can TEI encoding help to frame and model?

We propose to explore these questions in this follow-up to an earlier paper given by two of the current proposal's authors at the 2022 TEI conference, which reported on our initial work on our TEI-encoding of ApertureScience.com. This website, launched in 2006 and now non-functional, presented a work of interactive fiction implemented as a Flash simulation of a DOS-style user interface—all part of a trans-media storytelling strategy that connects the *Portal* and *Half-Life* video game franchises (Galey, 2023). In this presentation, we will

report on what we have learned from completing our critical edition of ApertureScience.com, which was recently published as a peer-reviewed micro-edition in the open-access journal *Scholarly Editing* (Galey, Forget and Allen, eds., 2025). If born-digital interactive fiction is new territory for TEI encoding, our paper can offer some early reconnaissance of a complex landscape that invites further exploration by the TEI community.

Informed by recent scholarship on electronic literature, digital ephemera, and born-digital textual artifacts (Kirschenbaum, 2008, 2021; Gibson, 2021; Pressman et al., 2015), on the idea of scholarly editions of born-digital works (O'Sullivan and Pidd, 2023), and on markup theory and the modelling of complex texts (Flanders and Jannidis, eds., 2018), our presentation will approach the questions above through specific examples of encoding decisions that we debated as a team. Specific examples will include: encoding artifacts of digital interactivity (e.g. command prompts, cursors, and animated text); modelling multi-linear reading paths and randomized text; and resolving, through editorial intervention, discrepancies between the text's decompiled Flash code and the edition's interface. Our encoding is transformed into a heavily customized Twine (<https://twinery.org>) interface, which includes unpublished textual material recovered from Flash code. Together, our examples will illuminate a perennial question in TEI encoding: what is the text, and what is the interface?

References

- Flanders, J. and Jannidis, F. (eds.) (2018) *The shape of data in the digital humanities: modeling texts and text-based resources*. New York: Routledge.
- Galey, A. (2023) 'Behind the scenes at ApertureScience.com: *Portal* and its paratexts', *Games and Culture*, 18(4), pp. 498-523.

- Galey, A., Forget, E. and Allen, B. (eds.) (2025) 'ApertureScience.com: a critical edition', *Scholarly Editing*, 42 [online]. Available at: <https://scholarlyediting.org/issues/42/aperturescience.com/>
- Gibson, R.H. (2021) *Paper electronic literature: an archaeology of born-digital materials*. Amherst, MA: University of Massachusetts Press.
- Kirschenbaum, M.K. (2008) *Mechanisms: new media and the forensic imagination*. Cambridge, MA: MIT Press.
- Kirschenbaum, M.K. (2021) *Bitstreams: the future of digital literary heritage*. Philadelphia: University of Pennsylvania Press.
- O'Sullivan, J. and Pidd, M. (2023) 'The born-digital in future scholarly editing and publishing', *Humanities and Social Sciences Communications*, 10(930) [online]. Available at: <https://doi.org/10.1057/s41599-023-02454-8>.
- Pressman, J., Marino, M.C., Douglass, J. (2015) *Reading project: a collaborative analysis of William Poundstone's Project for Tachistoscope (Bottomless Pit)*. Iowa City: University of Iowa Press.

Encoding Ambiguity

A TEI Approach to Argentinean Reading Theory in an 80s Literary Magazine

Federico Gabriel Cortés

Bergische Wuppertal
Universität, Germany

Keywords

Literary theory; TEI; Sitio; Argentinean intellectual history

Abstract

This presentation describes the ongoing creation of a TEI-based scholarly digital edition of SITIO, an important Argentine literary and cultural journal published during the tumultuous 1980s, from the late dictatorship to the Malvinas War. The magazine features a rich mix of genres—essays, literary criticism, fiction, poetry, legal texts, translations—by various authors, collectively engaging with a period of intense political and cultural upheaval.

Our central hypothesis identifies an implicit "Argentinean reading theory" within the magazine, characterized by a deliberate avoidance of interpretive 'overdetermination'. We argue the publication promotes the construction of multiple, ambiguous meanings as a critical strategy to challenge the stability of hegemonic discourse. This resistance is performed stylistically through pervasive irony and the frequent deployment of rhetorical devices, including metaphor, anacoluthon, paradox, and antithesis.

The primary encoding challenge, therefore, lies in applying TEI XML—often associated with structuring information and fixing meaning—to capture the conceptual landscape of texts defined by semantic ambiguity and resistance to singular interpretation. How can TEI’s tools for conceptual coding engage with this fluidity without imposing undue stability?

This paper explores strategies using TEI to navigate this tension. We discuss the use of elements like `<term>` and `/<seg>` (with `@type` or `@ana`) to identify key concepts and rhetorical figures. Critically, we examine the application of `<interpGrp>/@ana` for thematic tagging alongside layered `<note>` annotations (`@type`, `@resp`, `@target`) to articulate nuanced, even potentially contradictory, interpretations that reflect, rather than resolve, the texts’ inherent ambiguities. This project serves as a case study in adapting TEI methodologies to encode complex theoretical discourse and interpretive indeterminacy within historically significant texts.

Encoding Argentinian Eighteenth and Nineteenth-Century Drama with TEI

Gimena del Rio Riande (1);
Ulrike Henny-Krahmer (2)

(1) CONICET;
(2) University of Rostock

Keywords

TEI, drama

Abstract

This poster aims to explain the first steps of the TEI-XML encoding of ArDraCor, Argentinian Drama Corpus, the first DraCor (<https://dracor.org/>) corpus with Latin American texts.

ArDraCor is a joint effort between the HD LAB (Argentina) and RosDH (Germany) that has started digitizing and encoding in TEI a corpus of plays – *loas*, *cielitos*, *sainetes*, comedies, dramas, *zarzuelas*—written and performed between the eighteenth and nineteenth centuries in the territory that today corresponds to Argentina.

The first goal of our project was to survey previous studies and editions of our corpus and either digitize or set up a workflow that includes text recognition using OCR and converting existing OCRed text or other formats into TEI. Currently, we are using the OCR4all tool (Reul et al. 2019, see also Dennerlein et al. 2025) and an encoding in the EzDrama format (Skorinkin et al. 2022) to prepare the digital full texts for encoding in TEI.

So far, almost 200 drama texts have been identified and can be accessed on the metadata level at: <https://hdlab.space/te-atrar/>. The texts we have already encoded in XML-TEI can be accessed at <https://github.com/dracor-org/ardracor>. In the encoding, we put a special emphasis on genre assignments in the TEI header and, where necessary, create new data records for subgenres of Argentinian drama in Wikidata (see example 1). We also encode foreign language expressions in the text itself, as these frequently occur in the texts (see example 2).

```
<textClass>
  <keywords>
    <term source="wikidata"
      type="genreTitle">Sainete</term>
    <term source="#seibel">Sainete urbano</term>
  </keywords>
  <classCode scheme="http://www.wikidata.org/
    entity/">Q835067</classCode>
</textClass>
```

Example 1: Encoding of a subgenre assignment.

```
<sp who="#Carlos">
  <speaker>C  rlos</speaker>
  <stage><emph>, al p  blico.</emph></stage>
  <p>   Se  ores, ten  a yo raz  n? Es indudable:
  <foreign xml:lang="lat"><emph>Similia similibus,
    curantur.</emph></foreign></p>
</sp>
```

Example 2: Encoding of a passage of text in foreign language.

References

- Dennerlein, Katrin, Martin Rupnig, and Christian Reul. 2025. "Zum Aufbau digitaler Dramenkorpora. OCR4alltoDraCorTEI als Baustein f  r die Edition von maschinenlesbaren Versionen historischer Dramendrucke." In *DHd2025. Under Construction. Book of Abstracts*. Zenodo. <https://doi.org/10.5281/zenodo.14942992>.
- Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas B  ttner, and Frank Puppe. 2019. "OCR4all – An open-source tool providing a (semi-) automatic OCR workflow for historical printings." *Applied Sciences* 9 (22). <https://doi.org/10.3390/app9224853>.
- Skorinkin, Daniil. 2024. "EasyDrama: a lightweight solution for encoding plays in TEI/XML." In: *TEI 2024 Book of Abstracts*. Buenos Aires: Universidad del Salvador. <https://doi.org/10.5281/zenodo.13883242>.

Encoding Complexity

TEI Modeling in the “Forschungsportal BACH” Project

Nadine Quenouille

Sächsische Akademie der
Wissenschaften zu Leipzig,
Germany

Keywords

Forschungsportal BACH, Historical Texts
Annotation, Heterogeneous Corpus

Abstract

The long-term project „Forschungsportal BACH“ launched in 2023, is a collaborative effort between the Saxon Academy of Sciences and Humanities in Leipzig and the Bach Archive Leipzig. The goal of the project is to document, digitally process, and make available all extant non-musical documents from the family of Johann Sebastian Bach, spanning from the late 16th to the early 19th century, in an online research portal.

The textual sources include private and official correspondence, as well as educational and professional records, legal documents, and other related materials – such as student registers, timetables, account books, wills, petitions, official records, as well as copies and transcripts of various public documents. Letters represent only a small portion of this overall heterogeneous corpus.

In the project we established a complex workflow starting with visiting the archives and digitizing the sources, via their recording in the project’s database, automatic text recognition and transcription by “Transkribus”, the correction and

structural annotation of the text that takes place there, to the automatic conversion from Transkribus output in PAGE XML to TEI via XSLT and further textual annotations in “TEI Publisher”.

To ensure standardization and long-term usability, all texts are encoded in TEI-P5 format. This encoding follows established guidelines, although the diversity and specificity of the sources occasionally present challenges.

This presentation addresses specific challenges in modeling a structurally and semantically heterogeneous corpus, using three very different document types as examples – a will, a school register, and a coherent document written by a single scribe that contains transcripts and attachments of multiple documents. The focus is on issues of structuring, annotation, and semantic interpretation, especially in areas where the current TEI modules may not fully address the needs of the sources.

Furthermore, solutions will be proposed for how these challenges can be addressed within the current TEI framework – for example, through a nuanced combination of existing elements and careful modeling that closely aligns with the materiality of the sources. The aim is to make transparent the strategies developed within the project and to stimulate further discussion on how to approach structurally and semantically diverse sources, particularly in cases where the scalability of existing modules may not fully meet the demands of the project. Throughout, the encoding remains TEI-valid and is intentionally free of project-specific customizations in the form of a separate schema against which it could be validated, as the full scope of editorial requirements will only become clear during the course of the work. This presentation offers insights into the editorial practice of working with a heterogeneous corpus and aims to foster a discussion on modeling strategies for complex and diverse archival collections.

Encoding Hyperfiction

Preliminary Considerations for a Digital Edition of Susanne Berkenheger's "Hilfe! Ein Hypertext aus vier Kehlen" (1998)

Ulrike Henny-Krahmer

Universität Rostock, Germany

Keywords

hyperfiction, digital literature, Susanne
Berkenheger, computer-mediated
communication, digital edition

Abstract

In the early phase of the Internet, numerous works of genuinely digital literature were created that can be classified as *hyperfiction*. They use the technique of hypertext to link many text sections (so-called "lexia", Landow 1992) in a non-linear way via link structures. Readers interact with the texts by choosing specific click paths, making it possible to read a work in a variety of ways (Bajohr and Roloff 2024, Short 2024).

Such works have also been produced in German-speaking countries, including by the author Susanne Berkenheger (born 1963). One of her works, which was awarded the prize for the best work by an author on the Internet at the Internet literature competition in Ettlingen in 1999, is "Hilfe! Ein Hypertext aus vier Kehlen" (Berkenheger 1998), which tells the story of Jo, who is thrown off the plane into the sea or the mountains and who

meets the characters Ed, Pia, Lea, and Max, who have their various hopes regarding their relationship to Jo. The work uses linked HTML pages, Java Script elements and pop-up windows, whereby statements about the figures are made in separate windows.

"Hilfe!" has already been archived in a project of the German Literature Archive in Marbach (DLA Marbach 2015) in order to be preserved in principle, but how could it be sustainably represented and made accessible as a text object? Recently, as part of the TEI Guidelines, there is a chapter on computer-mediated communication, which covers "all kinds of communications that are mediated by digital technologies (such as text on web pages, written exchanges in chats and forums, interactions with artificial intelligence systems, the spoken conversations in internet video meetings)" (TEI Consortium 2025). Can the elements and attributes proposed in the CMC chapter be meaningfully used to encode hyperfiction or does this require further additions to the TEI Guidelines?

In this short presentation, initial thoughts are given on how the work "Hilfe!", as an example of hyperfiction from the 1990s, could be encoded in TEI. This involves an initial approach to the work from the user or reader's perspective, i.e. via the output in the browser. However, further considerations on representation, which also include source code and executed code, will certainly have to be made.

References

- Bajohr, Hannes and Simon Roloff. 2024. *Digitale Literatur: Zur Einführung*. Hamburg: Junius Verlag.
- Berkenheger, Susanne. 1998. "Hilfe! Ein Hypertext aus vier Kehlen." <http://www.wargla.de/hilfe.htm>. Accessed April 29, 2025.

DLA Marbach, ed. 2015. "Hilfe!" *Wiki des Projekts ,Netzliteratur authentisch archivieren und verfügbar machen.'* Last changed Mai 3, 2016, accessed April 29, 2025. <https://www.wik.dla-marbach.de/line/index.php?title=Hilfe!&oldid=4070>.

Landow, George P. 1992. *Hypertext. The Convergence of Contemporary Critical Theory and Technology*. Baltimore: Johns Hopkins University Press.

Short, Emily. 2024. "Narrative and Interactivity." In: *The Cambridge Companion to Literature in the Digital Age*, edited by Adam Hammond, 177-193. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009349567.012>.

TEI Consortium. 2025. "Computer-mediated Communication." In: *TEI: Guidelines for Electronic Text Encoding and Interchange*. P5 Version 4.9.0, 24th January. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CMC.html>.

Encoding language diversity in TEI

A description of regional and non-standard languages in multilingual diachronic corpora

Juliette Janès; Rachel Bawden;
Thibault Clérice; Rasul Dent;
Lucence Ing; Oriane Nédey;
Benoît Sagot

INRIA, Paris, France

Keywords

TEI ODD, fine-grained language description, non-standard languages, multilingual corpora

Abstract

Currently, French minority and regional languages lack digital resources, especially for cultural preservation and equitable representation. With the emergence of new tools for speakers, it is crucial to build stable corpora for under-resourced languages. In France, initiatives like DIVITAL⁹ and the earlier RESTAURE project¹⁰ have developed resources for specific languages, but much remains to be done.

⁹ <https://divital.gitpages.huma-num.fr/fr/>

¹⁰ <https://restaure.unistra.fr/>

To address this gap, the COLaF¹¹ project aims to develop corpora and Natural Language Processing (NLP) tools for the languages of France, spanning different languages and dialects, time periods, geographic regions and social contexts. Even within one language, variation happens depending on temporal, geographical and sociolinguistic factors (Labov, 1989), especially for non-standard languages. While NLP increasingly addresses linguistic diversity (Broeder, 2006) and TEI elements exist for this purpose, to our knowledge, no previous project has tackled the issue of a standardized structure for linguistic variation and its associated parameters in TEI corpora.

We propose a TEI schema to enable the fine-grained description of language variation, with a particular focus on non-standard languages.^{12,13} We use a combination of `<language>`, `<particDesc>` and `<settingDesc>` elements to document the manifestation of a language as it is spoken in a specific place, at a specific time and within a particular context. As BCP47 does not include all the ‘regional languages’ we would like to work on, we use Glottolog identifiers (Nordhoff, 2011), for a wider range of languages. We will later propose various language classifications to browse the corpus to address the multiple representations of languages and dialects. We specify both elements to include detailed metadata on the language, individuals associated with the document, and collection location, using controlled vocabularies where possible¹⁴ (Figures 1, 2 and 3).

11 COLaF (‘Corpus et Outils pour les Langues de France’ ‘Corpus and Tools for the Languages of France’), is an ongoing project funded by Inria to contribute to the development of free corpora and tools for French and other languages of France. For further information see: <https://colaf.huma-num.fr/en/>

12 The ODD schema and document is available here: <https://github.com/DEFI-COLaF/metadata>

13 The ODD has been already used for various multi-speaker corpora such as the Molye (<https://github.com/DEFI-COLaF/Molye>), Concours Picard (https://github.com/DEFI-COLaF/Datasets_text/tree/main/Picard_Concours) and Forum Occitania corpora.

14 For example, Glottolog identifiers for the language id, Geonames identifiers (<https://www.geonames.org/>) for the place id, etc.

The metadata is linked to the text via attributes `@who`, `@corresp`, `@hand`, and `@xml:lang` (Figure 4).

```
<language ident="lorr">
  <idno type="langue">lorr1242</idno>
  <idno type="script">latn</idno>
  <name>Lorrain Roman</name>
</language>
```

Fig. 1. Example of a content for a `<language>` element

```
<person xml:id="person_8">
  <name>Martin Bouchy</name>
  <langKnowledge>
    <langKnown level="maternal" tag="lorr"/>
    <langKnown level="fluent" tag="stan"/>
  </langKnowledge>
  <birth>
    <date when="1775"/>
    <placeName>
      <settlement>
        <name>Onville</name>
        <idno type="geonames">2989497</idno>
      </settlement>
      <region>
        <name>Meurthe et Moselle</name>
        <idno type="geonames">2994111</idno>
      </region>
    </placeName>
  </birth>
  <residence>
    <settlement>
      <name>Onville</name>
      <idno type="geonames">2989497</idno>
    </settlement>
    <region>
      <name>Meurthe et Moselle</name>
      <idno type="geonames">2994111</idno>
    </region>
  </residence>
</person>
```

Fig. 2. Example of a content for a `<person>` element

```

<place xml:id="place_11" corresp="#lorr">
  <settlement>
    <name>Onville</name>
    <idno type="geonames">2989497</idno>
  </settlement>
  <region>
    <name>Meurthe et Moselle</name>
    <idno type="geonames">2994111</idno>
  </region>
</place>

```

Fig. 3. Example of a content for a `<place>` element

```

<div type="parabole" n="12" hand="#person_8" xml:lang="lorr" corresp="#place_11">
  <head xml:lang="stan"><lb/>Traduction de la Parabole de l'Enfant Prodigue, en Patois d'Onvil
    de Gorze, département de la Moselle, envoyée par<lb/>M. Bouchy, d'Onville.</head>

```

Fig. 4. Example of a linked content

[2025-01-24]. <https://tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html#HD41> (2025-04-07)

TEI Consortium, eds. "16.2 Contextual Description" TEI P5: Guidelines for Electronic Text Encoding and Interchange. [2025-01-24]. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAH> (2025-04-07)

References

- Broeder, Daan, et Peter Wittenburg. « The IMDI metadata framework, its current application and future direction ». *Int. J. Metadata Semant. Ontologies* 1, no 2 (1 octobre 2006): 11932. <https://doi.org/10.1504/IJMSO.2006.011008>.
- Labov, William. *Sociolinguistique*. Traduit par Alain Kihm. 1 vol. Le sens commun. Paris: Minuit, 1989.
- Nordhoff, Sebastian, et Harald Hammarström. « Glottolog/Langdoc: Defining Dialects, Languages and Language Families as Collections of Resources ». In *Proceedings of the First International Workshop on Linked Science 2011*, 783:7. Bonn, Germany: CEUR, WS, 2011. <https://ceur-ws.org/Vol-783/paper7.pdf>.
- TEI Consortium, eds. "2.4.2. Language Usage" TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Encoding Ligature Glyphs in Arabic Manuscripts Using TEI-XML

Serhat Acar

UBFC, Germany

Keywords

Arabic script encoding, Ligatures in Arabic script, TEI customization Arabic

Abstract

Encoding right-to-left scripts in TEI-XML, specifically Arabic, presents significant technical challenges, particularly in representing ligature glyphs and complex character components. Arabic script is characterized by its use of numerous ligatures—combinations of characters that are written as a single, unified glyph. Properly encoding these ligatures is essential for maintaining the integrity and readability of the manuscripts.

A primary issue is that standard TEI-XML tags and attributes may not sufficiently capture the nuances of these ligatures. The TEI framework provides the `<g>` element to specify individual glyphs or character components, ensuring accurate representation and searchability. For instance, an Arabic ligature such as “J” (lam-alif) can be encoded as:

```
<p><g ref="#ligature_arabic_lam_alef"> ال </g> م ادخست انكم ي  
ي برع ال ص ن ال  
ل ص ت م ال . </p>
```

This approach requires a detailed markup scheme, and a comprehensive understanding of the specific ligatures used in the

manuscript. This can complicate the encoding process, especially for editors who may not be familiar with all the nuances of Arabic script. Accurately representing these ligatures involves ensuring that the TEI-XML documents maintain the correct rendering of text flow and character combination, as managed by the Unicode Bidirectional Algorithm. This is crucial for preserving the original format and readability of the manuscripts. For example, managing punctuation and embedded Latin text within Arabic script requires careful handling to ensure proper display and interpretation:

```
<pb n="3" facs="arabic-script-page3"> <bidirOverride  
dir="rtl">  
<p> <hi rend="italic">Latin</hi> ن ي م ض ت ع م ي ب ر ع ال ص ن ال ي ل ع ل ا ث م  
ص ن . </p>  
</bidirOverride> </pb>
```

This proposal aims, to focus on the challenges and solutions related to encoding ligature glyphs and complex character components in Arabic manuscripts using TEI-XML. By leveraging the TEI framework's capabilities, it is possible to create a robust encoding scheme that faithfully represents Arabic ligatures, enhancing the accessibility of these texts for scholarly research.

References

- Unicode Consortium. (2021). Unicode Bidirectional Algorithm. Retrieved from <https://www.unicode.org/reports/tr9/tr9-46.html>
- TEI Consortium. (n.d.). TEI Guidelines. Retrieved from <http://www.tei-c.org/Guidelines/>

Encoding Multilingual and Multiscript Sources in TEI

A historical Spanish-Chinese dictionary

Martina Scholger; Elisabeth Steiner;
Sabrina Strutz; Melanie Frauendorfer

University of Graz,
Austria

Keywords

multilingual, multiscript, digital scholarly
edition, historical lexicography, tonal
languages

Abstract

The importance of documenting yet under-resourced language varieties and strengthening the collaboration between digital philology and language documentation has gained increasing recognition (Bowers, 2020; Czaykowska-Higgins et al., 2014; Ngué Um, 2017; Thompson, 2023; VedaWeb, 2024). In this context, we present our approach to encoding a multilingual, multiscript historical dictionary using the TEI standard. We focus on addressing key challenges, including orthographic variation, the representation of tonal features, and the need to balance historical fidelity with the requirements of contemporary research.

The Early Manila Hokkien (EMHo, Döhla et al., 2024, Döhla et al., 2022) project digitizes and analyzes the 17th-century “Bocabulario de lengua sangleya por las letraz de el A.B.C.” (henceforth Bocabulario, British Library Add ms. 25.317), a Chinese-Spanish dictionary documenting the Hokkien variant

spoken by Chinese immigrants in early Manila. It serves as a valuable witness to historical language contact phenomena and provides unique insights into the massive linguistic and cultural exchange involving multiple languages and ethnic groups during this period in the Philippines, in particular the Spanish-Chinese relations (Döhla, 2025).

A fundamental tension in our multi-layered encoding approach lies between diplomatic accuracy to the historical source, and the creation of structured research data that meets modern scholarly requirements and computational tractability. This ‘representational gap’ manifests in several ways. For example, the Bocabulario contains inconsistent orthography and spacing, requiring careful decisions about the depth of normalization.

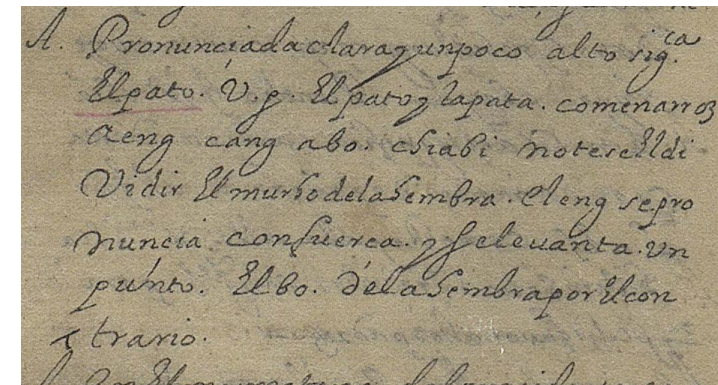


Fig. 1: Add ms. 25.317, fol. 2v. Facsimile reused with kind permission © British Library Board, all rights reserved.

The data model is based on the Lex-O TEI dictionary encoding recommendations (Tasovac et al., 2018). A simple dictionary entry (Fig. 1) includes a lemma (<form>), a word sense group (<sense>), a definition (<def>), and examples (<cit>) in Hokkien, along with Spanish translations. We use <entryFree> to preserve the original spelling, punctuation, and structure for diplomatic transcription,

while editorial interventions and notes enable the creation of a normalized and critical edition. The `<entry>` element expands the structure beyond the historical source by incorporating expert-supplied modern Taiwanese romanization, Chinese characters, and English translations.

The markup addresses the multilingual and multiscript challenges by applying the `@xml:lang` attribute to identify language varieties and their script representations, following the IETF BCP 47 standard (Phillipps and Davis, 2009). The constant switching requires meticulous attention to markup consistency.

In addition, we implement a linguistic analysis layer that integrates interlinear glossing following the Leipzig Glossing Rules (EVA MPG, 2015). The project aims to contribute methodologically to the TEI Guidelines by developing a recommendation for encoding tonal features which occur in approximately 70% of the world's languages (Yip, 2002; Maddieson, 2013).

We particularly emphasize interoperability to enable cross-domain reuse of this unique historical source by both linguists and historians. Thus, we aim to bridge the gap between multiple disciplines and address the challenges arising from that. Furthermore, this work contributes to broader efforts to document multilingual language contact scenarios. _

References

- Bowers, J. (2020) 'Language documentation and standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec'. *Computation and Language [cs.CL]*. École Pratique des Hautes Études. <https://tel.archives-ouvertes.fr/tel-03131936>.
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J. and Hudson, M. (2020). 'The CARE Principles

```
<entryFree type="original" xml:id="entry.0005" facs="#IMG.002v">
  <form type="lemma" xml:lang="nan-Latn-PH">A</form>
  <sense xml:lang="es-Latn-PH">
    Pronunciada clara<supplied> </supplied>y<supplied>un
    <supplied> </supplied>poco alto<supplied>;</supplied>
    <choice>
      <abbr>sig<am>_</am><hi rend="superscript">ca</hi></abbr>
      <expan>sig<ex>_nifi</ex>ca</expan>
    </choice>
    <lb/>
    <def>El <supplied> </supplied>gato</def><supplied>;</supplied>
    <choice>
      <orig><choice><abbr>V.g.</abbr><expan>verbi gratia</expan></choice></orig>
      <reg><choice><abbr>p.ej.</abbr><expan>por ejemplo</expan></choice></reg>
    </choice>
    <cit type="example" xml:lang="es-Latn-PH" xml:id="entry.0005.ex.1">
      <quote>El gato y la<supplied> </supplied>gata<surplus></surplus>
      <supplied> </supplied>arroz</quote><lb/>
      <cit type="translation" xml:lang="nan-Latn-PH">
        <quote>a<supplied> </supplied>gng gang a<supplied> </supplied>
        <supplied> </supplied>bia<supplied> </supplied>bi</quote>
      </cit>
    </cit><supplied> </supplied>
    <choice><orig></orig><reg>ó</reg></choice>tese<supplied> </supplied>gl
    di<lb break="no"/>Vidir. El m<choice><orig>ur</orig><reg>ac</reg>
    </choice>ho<note n="1"><p>[...]</p></note> de la
    hembra<surplus>_</surplus><supplied> </supplied>
    [...]
```

Code Example 1: Diplomatic transcription, critical edition, and translation of one dictionary entry. Simplified and shortened.

for Indigenous Data Governance'. *Data Science Journal*, 19: XX, pp. 1-12. DOI: <https://doi.org/10.5334/dsj-2020-043>.

Ozaykowska-Higgins, E., Holmes, M. D. and Kell, S. M. (2014) 'Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxoin Database-Dictionary Project'. *Language Documentation & Conservation* 8: 1-37.

Döhla, H. (2025) 'The Bocabulario de la Lengua Sangleya por las Letras de el A.B.C. (Manila, ca. 1617): structure, contents, perspectives'. *Historiographia Linguistica*. <https://doi.org/10.1075/hl.00170.doh>.

Döhla, H., Klöter, H. and Scholger, S. (eds.). (2024) 'Early Manila Hokkien. Digital Edition of the *Bocabulario de lengua sangleya*'. GAMS. <https://gams.uni-graz.at/emho>.

Döhla, H., Klöter, H., Scholger, M. and Steiner, E. (2022) 'Annotating a historical manuscript as a linguistic resource'. *TEI2022 Conference Book* (1.1). TEI2022 - "Text as Data" (TEI2022), edited by James Cummings. Newcastle, UK. Zenodo. <https://doi.org/10.5281/zenodo.7120027>, p. 129.

EVA MPG (= Max Planck Institute for Evolutionary Anthropology, Department of Linguistics). (2015). 'The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses'. <https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf>.

Maddieson, I. (2013) 'Tone', in Dryer, M. S. and Haspelmath, M. (eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/13>.

Ngué Um, E. (2017) 'Issues in digital text representation, online dissemination, sharing and reuse for African tone languages'. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu, Hawai'i, 24-32. <https://aclanthology.org/W17-0104.pdf>.

Phillips, A. and Davis, M. (eds.). (2009) 'Tags for Identifying Languages', *BCP 47, RFC 5646*, DOI 10.17487/RFC5646, <https://www.rfc-editor.org/info/rfc5646>.

Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. and Witt, A. (2018) 'TEI Lex-O: A baseline encoding for lexicographic data'. Version 0.9.3. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILexO/TEILexO.html>.

Thompson, W. (2023) *Epifanii Slavinskii's Greek-Slavonic-Latin Lexicon Between East and West*. PhD Thesis, Heidelberg.

TEI Consortium (eds.). (2025) 'IO Dictionaries'. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.8.1. TEI Consortium. <https://tei-c.org/Vault/P5/4.8.1/doc/tei-p5-doc/en/html/DI.html>.

VedaWeb. (2024) 'Online Research Platform for Old Indic Texts'. <https://vedaweb.uni-koeln.de>.

Yip, M. (2002). *Tone*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139164559>.

Enriching textual data from digital libraries

Boris Lehečka

Moravian Library in Brno, Czech Republic

Keywords

digital library, natural language processing, text enrichment

Abstract

The demonstration will introduce open-source tools that can mine and enrich data from Czech digital libraries that use the Kramerius system.

In 2019-2022, the DL4DH project (Development of tools for more effective use and mining of digital library data to strengthen digital humanities research) has developed a set of applications, DL4DH Kramerius+, Feeder and TEI Converter, that allow researchers to retrieve data and metadata from digital library not only in existing original formats (FOXML, MODS, ALTO, plain text, JPG), but also in TEI format, which can contain morphological data and recognized entities in addition to the text itself. Data are enriched by external tools, UDPipe and NameTag and transformed into the TEI format. Based on a contract between the library and the research institution (universities, institutes of the Academy of Sciences, etc.), researchers can request the enrichment of specific titles and download them for research, e.g. for training artificial intelligence.

The demonstration will also include the Libri augmentati software application. Like the server applications from the DL4DH project, it serves the same purpose, but it is programmed in

XProc 3.0 and can run on desktop computers running Windows, macOS or Linux. It is intended for individual researchers and for processing smaller volumes of publications, especially freely available (in Czech digital libraries monographs published before 1950, periodicals before 1910). The application comprises a set of modules: for enrichment text from ALTO format or plain text using external tools, UDPipe and NameTag; all data are at the end converted to TEI format. The modules can be developed to be more generic so that they can be used not only with Kramerius digital libraries but also with digital libraries that use different APIs and metadata, such as IIIF.

References

- Knihovna Akademie věd ČR, Národní knihovna ČR, Moravská zemská knihovna v Brně and InQool, (2022a). DL4DH Kramerius+ [online]. Version 0.10.0. [computer software]. [Viewed 11 July 2025]. Available from: <https://github.com/LIBCAS/DL4DH-Kramerius-plus>
- Knihovna Akademie věd ČR, Národní knihovna ČR, Moravská zemská knihovna v Brně and InQool, (2022b). DL4DH TEI Converter [online]. Version 0.6. [computer software]. [Viewed 11 July 2025]. Available from: <https://github.com/LIBCAS/DL4DH-TEI-Converter>
- Knihovna Akademie věd ČR, Národní knihovna ČR, Moravská zemská knihovna v Brně and InQool, (2023). DL4DH Feeder [online]. Version 1.3.0. [computer software]. [Viewed 11 July 2025]. Available from: <https://github.com/LIBCAS/DL4DH-Feeder>
- Boris Lehečka, (2025). Libri augmentati [online]. Version 1.0.0. [computer software]. Brno: Moravian Library in Brno. [Viewed 11 July 2025]. Available from: <https://github.com/moravianlibrary/libri-augmentati>

Lehečka, B., Novák, D., Kersch, F., Hladík, R., Bišková, J., Sekyrová, K., Válek, F., Vozár, Z., Bodnár, N., Sekan, P., Bežová, M., Žabička, P., Lhoták, M. and Straňák, P., (2022). *Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách* [online]. Knihovna AV ČR, v. v. i. [Viewed 11 July 2025]. Available from: <https://hdl.handle.net/11104/0335692>

Straka, M. and Straková, J., (2025). NameTag [online]. Version 3.1.0. [computer software]. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy. [Viewed 11 July 2025]. Available from: <http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E>

Straka, M. and Straková, J., (2023). UDPipe [online]. Version 2.1.0. [computer software]. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy. [Viewed 11 July 2025]. Available from: <http://hdl.handle.net/11234/1-1702>

Exploring correspSearch v3 as a Service for Minimal Editions of Correspondence

Stefan Dumont (1);
Sascha Grabsch (1);
Jonas Müller-Laackman (2);
Ruth Sander (1);
Steven Sobkowski (1)

(1) Berlin-Brandenburg
Academy of Sciences and
Humanities, Germany;
(2) State and University
Hamburg Carl von Ossietzky

Keywords

correspondence, minimal edition, sustainability, web service, interoperability

Abstract

Since 2014, correspSearch aggregates the metadata of edited letters provided in the *Correspondence Metadata Interchange Format* (CMIF, developed by TEI Correspondence SIG) from different projects. With version 3 of correspSearch, a full-text search in the edited letters is now possible for the first time. To enable this, a URL to the corresponding TEI-XML file must be specified for each letter in the CMIF file, which correspSearch then retrieves. The TEI structure is also taken into account: letter text, editor's comments and greeting/closing formulae can be distinguished. Additionally correspSearch v3 offers also different visualisations of search results or individual CMIF files: timeline of the correspondence(s), geographical distribution of sending and receiving places and a network visualisation.

Furthermore, it has recently become possible to search and filter for persons *mentioned* in the letters.

The new functions also support the editing projects themselves. For example, the visualisations offer a graphical overview of the project's data, which is often not available in the project itself. Some edition projects have therefore integrated these visualisations into their own digital offerings. One example of this is the Graz Nunciature Reports.¹⁵

The *Buber Korrespondenzen digital*¹⁶ goes a step further: The project, which runs until 2045, does not yet have a web platform, although letters edited in TEI-XML are already available. The project currently publishes these in a Gitlab repository. With the help of TEI Boilerplate, an HTML view is available on Gitlab pages.¹⁷ This web presentation is referenced from the CMIF metadata. As a result, there is a human-readable view for each individual letter, but access in the sense of an overview with search and filter options is currently only offered via correspSearch.¹⁸

Both use cases lead to the question of whether digital editions of correspondence could be made more sustainable if correspSearch would be included as part of their conception. Common solutions for digital editions currently require a database and technologies based on it. However, these need to be maintained and updated even after the project has been completed. A more sustainable approach would be to publish the digital edition as a TEI-XML dataset and a static HTML website (generated via XSLT, TEI Boilerplate etc.). Both would require no further maintenance. The integration of correspSearch could

15 <https://grazer-nuntiatur.acdh.oeaw.ac.at/visualizations.html>

16 <http://buber-korrespondenzen.digital/>

17 The presentation view with TEI Boilerplate is currently temporarily deactivated because it is being adapted more closely to the needs of the project.

18 <https://correspsearch.net/de/suche.html?e=ae451a0a-2186-450c-ae5d-72024981937c&x=1&w=0>

replace the need for a database and dynamic programming by offering a full-text search with extensive filtering options as well as visualisations.

The prerequisites for this are already well laid out in correspSearch: The CMIF files as well as the TEI XML files for the full-text search, can be stored permanently and statically on a web space or in a repository without the need for dynamic APIs. It also does not matter which technology was used to create the TEI XML data or the static HTML.

The possibility of creating digital letter editions with manageable effort and without permanent maintenance obligations means that correspondence for which no larger DH resources can be provided could be published digitally. Often this concerns letters from non-canonised writers (such as women).

With this in mind, our lecture will present the outlined possibilities and also discuss different types of integration, including their advantages and disadvantages in the context of 'sustainability' and 'minimal computing / editions'. A specially created practical example will illustrate the interaction of correspSearch and static digital letter editions in the lecture.

References

- Calarco, Gabriel A. 2023. 'La éocfrasis en el Libro de Alexandre, un proyecto de edición digital para el estudio de la poesía clerical castellana del siglo XIII con minimal computing'. *Journal of the Text Encoding Initiative*, no. Issue 16 (May). <https://doi.org/10.4000/jtei.4494>.
- Crompton, Constance. 2023. "'No Boutique or Fashionable Technologies": Project Development, Mentorship, and Sustainability in an Innovation-First World'. *Digital Humanities Quarterly* 17 (1). <https://digitalhumanities.org/dhq/vol/17/1/000660/000660.html>.

Dombrowski, Quinn. 2022. 'Minimizing Computing Maximizes Labor'. *Digital Humanities Quarterly* 16 (2). <https://www.digital-humanities.org/dhq/vol/16/2/000594/000594.html>.

Dumont, Stefan. 2018. „correspSearch – Connecting Scholarly Editions of Letters“. *Journal of the Text Encoding Initiative* 10. <https://doi.org/10.4000/jtei.1742>.

Dumont, Stefan, Ingo Börner, Dominik Leipold, Jonas Müller-Laackman, und Gerlinde Schneider. 2019. „Correspondence Metadata Interchange Format“. In *Encoding Correspondence. A Manual for TEI-XML-Based Encoding of Letters and Postcards*, edited by Stefan Dumont, Susanne Haaf, und Sabine Seifert. Berlin. <https://encoding-correspondence.bbaw.de/v1/CMIF.html>.

Giannetti, Francesca. 2019. "'So near While Apart': Correspondence Editions as Critical Library Pedagogy and Digital Humanities Methodology". *The Journal of Academic Librarianship* 45 (5): 102033. <https://doi.org/10.1016/j.acalib.2019.05.001>.

TEI Correspondence SIG. 2025. „Correspondence Metadata Interchange Format (CMIF) v1.1.0“. <https://github.com/TEI-Correspondence-SIG/CMIF>.

Pierazzo, Elena. 2019. 'What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter'. *International Journal of Digital Humanities* 1 (2): 209-20. <https://doi.org/10.1007/s42803-019-00019-3>.

Stadler, Peter. 2014. 'Interoperabilität von Digitalen Briefeditionen'. In *Fontanes Briefe Ediert*, edited by Hanna Delf von Wolzhagen, 278-87. Fontaneana 12. Würzburg: Königshausen & Neumann.

Stadler, Peter, Marcel Illetschko, and Sabine Seifert. 2016. 'Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>'. *Journal of the Text Encoding Initiative [Online]* 9. <https://dx.doi.org/10.4000/jtei.1742>.

Viglianti, Raffaele, Gimena del Río Riande, Nidia Hernández, and Romina De León. 2022. 'Open, Equitable, and Minimal: Teaching Digital Scholarly Editing North and South'. *Digital Humanities Quarterly* 16 (2). <https://www.digitalhumanities.org/dhq/vol/16/2/000591/000591.html#>.

Eye Rhyme in the Digital Victorian Periodical Poetry Project

Martin David Holmes

University of Victoria, Canada

Keywords

rhyme, encoding poetry, digital edition

Abstract

In my presentation to the 2022 TEI Members Meeting, “Encoding sonic devices: what is it good for?”, I described the encoding of sonic devices, including rhyme, in the Digital Victorian Periodical Poetry (DVPP) project. That presentation made some cautious well-caveated suggestions of possible uses of this encoding, but ended with this assertion: “The most important thing is to surface this encoding for end-users and readers. Make it visible. Make it obvious. Make it usable.” Three years later, and further into the project, I’m now one of those end-users. I have become interested in forms of rhyme which are somehow deficient or unsatisfying to the reader, in particular „eye rhyme” (also known as “visual rhyme”, “graphic rhyme”, etc.: see Rickert). Two words are categorized as an eye-rhyme where it appears from their orthography that they should rhyme, but sonically they don’t. Common examples would be “cow” and “low” or “cough” and “through”. My interest is well-served by the DVPP collection; at the time of writing, we have tagged 1,335 instances of eye-rhyme, distributed across 745 encoded poems.

Eye-rhyme is particularly intriguing because it may arise from a number of distinct poetic intentions. Some eye-rhyme is

simply the resort of a bad or lazy poet, and during the Victorian period, this is how most scholars and arbiters of poetic taste viewed it (Mazel 2014, 131-133). It may also be a fossilized rhyme – a rhyme which in a previous era was a true rhyme, but has been distorted by sound change. Such rhymes are often used by later poets due to tradition. On the other side of the diachronic scale, a rhyme that was consonant when the poet wrote it may be an eye-rhyme to a modern reader; and even synchronically, a rhyme which is perfectly good in the dialect of the poet may appear as an eye-rhyme to a reader whose dialect is different. Finally, such rhymes may be purposefully deployed as a comic or satirical device. Thus eye-rhyme, and the perception and judgement of it, is at the intersection of poetic tradition, literary mores, historical sound-change, linguistic register and dialect, and poetic skill and intent.

This presentation will look at specific examples of eye-rhyme from Victorian periodicals, and examine its usage over time and across different publications, with their distinct audiences and orientations. I hope to challenge the instinctive attitude many native English speakers (including myself) share with the Victorians, that eye-rhyme is somehow displeasing, inadequate, or substandard, by looking at cases where it enhances the poem, or is consciously deployed for effect. I will also look at cases where the DVPP team has tagged eye-rhymes erroneously or controversially, drawing on the many discussions we have had over historical pronunciation and dialect; these discussions put us, as modern readers, in dialogue with the poetry, and this is one of the best reasons for encoding in the first place.

References

Chapman, Alison (ed.) and the DVPP team. 2025. *Digital Victorian Periodical Poetry Project*, Edition 0.98.10beta, University of Victoria. <https://dvpp.uvic.ca/>.

Holmes, Martin. 2022. "Encoding sonic devices: what is it good for?" Text Encoding Initiative Conference 2022, Newcastle, UK, September 14, 2022. <https://zenodo.org/record/7089666>.

Mazel, Adam Martin. 2014. The Work and Play of Rhyme in Victorian Verse Cultures, 1850-1900. Unpublished PhD Dissertation. <https://deepblue.lib.umich.edu/handle/2027.42/108906>.

Rickert, William E. 1978. "RHYME TERMS." *Style* 12, no. 1: 35-46. <https://www.jstor.org/stable/pdf/45109024.pdf>.

FAIR research data from Goethe's inbox: The first 2,400 incoming letters as TEI-XML full texts

On the publication of the "Letters
to Goethe" in the Academy Project
PROPYLÄEN:
Goethe's Biographica¹⁹

Christian Thomas (1,2);
Katharina Hofmann-Polster (1);
Claudia Häfner (1)"

(1) Klassik Stiftung
Weimar, Germany;
(2) Berlin-Brandenburg
Academy of Sciences
and Humanities, Germany

Keywords

digital scholarly edition, digital humanities,
corpus, letters, metadata, standardised
data, authority files, Goethe, TEI-XML, API,
correspSearch, CMIF, correspDesc

Contributor Roles, following the CRediT Taxonomy:

Christian Thomas (Conceptualization, Writing - original draft),
Claudia Häfner, Katharina Hofmann-Polster (Writing - review &
editing).

¹⁹ This abstract is a condensed, more TEI-centred version of our contribution to this year's conference of the 'Digital Humanities in the German-speaking Countries' association, *DHd 2025*; abstract available at <https://doi.org/10.5281/zenodo.14943015>.

Abstract

The large-scale Academy project *PROPYLÄEN: Goethes Biographica*, which is scheduled to run until 2039, combines four originally independent editions: Johann Wolfgang von Goethe's journals, his 'Encounters and Conversations', his outgoing letters, and finally, his incoming correspondence. Since September 2024, the project's edition and research platform <https://goethe-biographica.de/> provides a comprehensive, TEI-XML-encoded corpus of freely licensed research data for the first time. The data set comprises 2,404 letters to Goethe (i. e. the first batch of up to 20,000 letters to come), including structured metadata, indices and references. The corpus creation was assisted by Automatic Text Recognition (ATR) tools like *Transkribus*. The TEI-XML was enriched and linked with identifiers from authority files. The metadata and full texts were fully integrated into the *correspSearch* web service, by generating CMIF files from the relevant information in the *correspDesc* element in the TEI header.²⁰

While developing the edition's TEI data model and preparing the corpus texts, some phenomena could not yet be annotated within the TEI framework, e. g. the several cases where a postscript is actually *not* located at the end of the letter, etc. In cases like this, we proposed improvements to the TEI P5 Guidelines to the community, which were discussed in issues on the TEI's GitHub repository, and finally implemented in the latest release of the guidelines.²¹ Thus, the *PROPYLÄEN* project became visi-

²⁰ CMIF = "Correspondence Metadata Interchange Format".

Since we publish the metadata first, there are already 15,312 letters to Goethe available via *correspSearch*, including the 2,404 letters with abstract and full-text mentioned in this abstract; cf. <https://correspsearch.net/de/suche.html?c=https://zenodo.org/record/14998880/files/ra-cmif.xml>.

²¹ Cf. TEI-C/TEI GitHub, e. g. Issues [#2542](#), [#2550](#), [#2516](#), and [#2292](#), all resolved with [TEI P5 Release 4.9.0 \(2025-01-24\)](#).

ble both as a user of the TEI as well as a contributor to the TEI standard.

The *PROPYLÄEN* platform is continuously developed, updated and expanded with additional data sets from all sub-editions. The poster provides an overview of the TEI corpus and offers an outlook on the upcoming, much more extensive research data from the *PROPYLÄEN* project. The technological, methodological and editorial principles as well initial analyses of the data are presented.

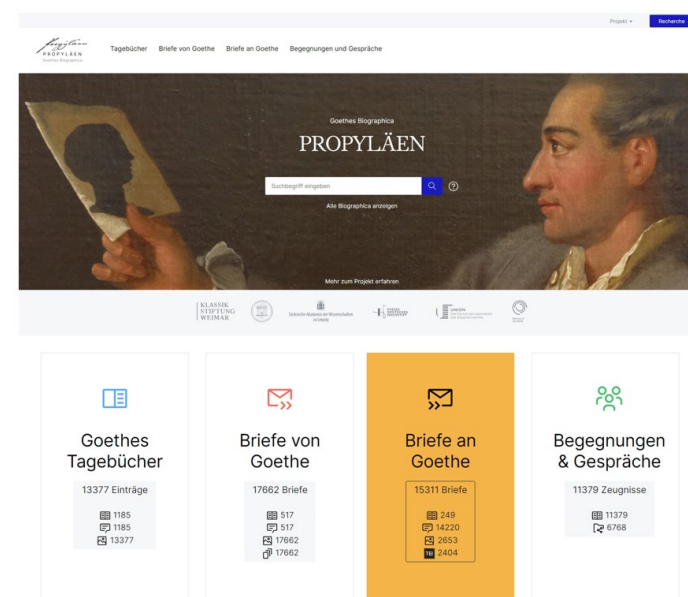


Fig. 1: PROPYLÄEN home page (detail) with information on the currently available data sets of the four subprojects or editions

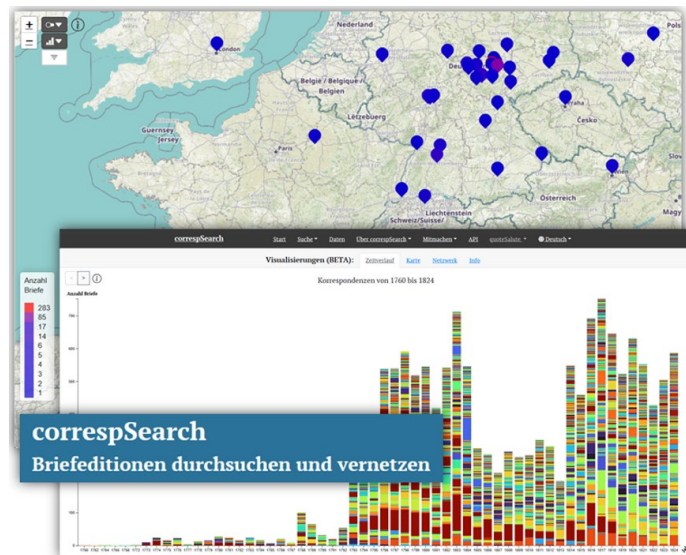


Fig. 4: Screenshots (detail) of additional visualisations via *correspSearch*: Map view of letters sent to Goethe in 1799; Timeline Overview of Correspondences in the current *PROPYLÄEN* dataset showing 15.312 letters from 1760 until 1824, URL <https://correspsearch.net/de/vis.html?e=ca477491-3e69-4332-977a-a25946837eb2&x=1&w=0&vistype=0>

Foregrounding Text Analysis in Digital Scholarly Editing

Who Are We Encoding For?

Diane Jakacki

Bucknell University, USA

Keywords

digital scholarly editing, text analysis, semantic encoding

Abstract

The goal of any scholarly textual endeavour, whether digital resource or analog press, is to produce texts. To maintain credibility and draw academic readers, the directors of these endeavours must establish rigorous production guidelines that are designed to drive editors to the publication finish line. When we, as scholarly textual editors, agree to undertake a play or poem or novel for such a publisher we do so with the understanding that we will adhere to their production and style guidelines, space limits, and defined editorial and critical apparatus. We accept that there will be constraints to our work so that it is consistent with other texts in the imprint and with the aim that it can be submitted in a timely manner. Many of us in the TEI community have worked for years in such spaces. But a tension remains for the digital scholarly editor: *when* can we undertake the philological, semantic, and contextual research that drew us

to the text in the first place? Must we wait until after we have completed all the required editorial pieces? How can we leverage our encoding to pursue our questions and integrate our textual analysis into the process without impeding the needs of our publisher? And how important is it for our encoding work to appear in the final published edition? Can we not encode for our own needs at the same time that we are encoding for our (imagined) readers?

In this talk I hope to propose that by building intentional phases for textual analysis into our editing process, we enrich the ultimate published edition, even if the published artifact doesn't 'show' our textual analysis. Whether it be 'seeing through' our text in *Voyant*, or shifting from code-centric *Oxygen* to the text-forward *LEAF-Writer environment*, or revealing the patterns in our tagging in tools like the *Dynamic Table of Contexts*, if we can take the time to examine our work while we are editing it from different vantage points I suggest that we can produce richer and more complex editorial work that what seems like an editorial luxury in fact makes us better editors.

I will use my work on Shakespeare and Fletcher's *Henry VIII* or *All is True* for the New Internet Shakespeare Editions²² to model different ways of intentionally integrating textual analysis into the encoding process, drawing on my own research questions about the play. My questions about royal identity, *Henry VIII* as a London City Play, and echoes in the play of Tudor propaganda for a playgoing audience in James I's England have changed radically as a result of the sidebar text analysis. Ultimately, I'd like to encourage a discussion about how we as editors can give ourselves the opportunity to pause the publishing process and dive into encoding leveraged analysis at the same time that we honour the commitments we have made to publish our work.

²² With gratitude for patience (and apologies for tardiness) to the general editors of the New Internet Shakespeare Editions, which was still the Internet Shakespeare Editions when I first submitted my edition proposal.

References

NB: these authors and collaborators have greatly informed my thinking in this way. While they do not necessarily appear by name in the abstract, their work is in my head and I hope to draw much of it out more explicitly in my paper.

- Brown, Susan, Brent Nelson, Stan Ruecker, Stéfan Sinclair, Nadin Adelaar, Ruth Knechtel, and Jennifer Windsor. 2013. "Text Encoding, the Index, and the Dynamic Table of Contexts." at the Annual Meeting of the Alliance of Digital Humanities. Lincoln, Nebraska.
- Cummings, James. 2019. "Opening the Book: Data Models and Distractions in Digital Scholarly Editing." *International Journal of Digital Humanities* 1 (2): 179–93. <https://doi.org/10.1007/s42803-019-00016-6>.
- Jakacki, Diane K. 2025. "Surfacing Encoding-Driven Analysis in Digital Editions with DToC." At the Workshop GREN-CRIHN « New Perspectives on Critical Editions » (Part 2). Montreal. <https://www.crihn.org/nouvelles/2025/03/13/workshop-new-perspectives-on-critical-editions-part-2/>
- . 2023. "Balancing Need, Speed, and the Future of Collaboration in Digital Scholarly Production." Invited Talk. Digital Humanities Virtual Seminar hosted by THINC Lab (U Guelph), CRIHN (UdeM), The Humanities Data Lab (UOttawa), and the Centre for DH (Toronto Metropolitan University).
- . 2018. "Internet Shakespeare Editions and the Infinite (Editorial) Others: Supporting Critical Tagsets for Linked Editions." In *Shakespeare's Language in Digital Media: Old Words, New Tools*, edited by Janelle Jenstad, Mark Kaethler, and Jennifer Roberts-Smith, 157–71. Routledge.
- . n.d. "Henry VIII: Internet Shakespeare Editions." Accessed April 27, 2025. <https://internetshakespeare.uvic.ca/Library/Texts/H8/>.

Jakacki, Diane K. and Susan Brown. 2024. "'I don't even see the code:' The Dynamic Table of Contexts - Optimizing our Encoded Texts for Web-based Reading and Analysis Environments." TEI 2024 Annual Meeting.

Jakacki, Diane K. and Janelle Jenstad. 2016. "Mapping Toponyms in Early Modern Plays with the Map of Early Modern London and Internet Shakespeare Editions Projects." In *Early Modern Studies and the Digital Turn*, edited by Laura Estill, Diane Jakacki, Michael Ullyot, 237-258. ITER.

Jenstad, Janelle. 2021. "Janelle Jenstad: Internet Shakespeare Editions - Speaking of Shakespeare." SoS #16. <https://www.buzzsprout.com/1732460/episodes/9123329-sos-16-janelle-jenstad-internet-shakespeare-editions>.

McGann, Jerome. 2014. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, Massachusetts: Harvard University Press.

-. 2001. *Radiant Textuality: Literature After the World Wide Web*. New York: St. Martin's Press.

Masten, Jeffrey. 2001. "More or Less: Editing the Collaborative." *Shakespeare Studies* 29:109-31.

Orley, Emily. 2022. "Editing as Creative Act: An Experiment in Speculative Thinking." *Textual Cultures* 15 (1): 11-17.

Pierazzo, Elena. 2016. "Modelling Digital Scholarly Editing: From Plato to Heraclitus." In *Digital Scholarly Editing*, edited by Elena Pierazzo and Matthew James Driscoll, 1st ed., 4:41-58. Theories and Practices. Open Book Publishers. <https://www.jstor.org/stable/j.ctt1fzhh6v.7>.

-. 2011. "A Rationale of Digital Documentary Editions." *Literary and Linguistic Computing* 26 (4): 463-77. <https://doi.org/10.1093/lc/fqr033>.

Rasmussen, Krista Stinne Greve. 2016. "Reading or Using a Digital Edition?: Reader Roles in Scholarly Editions." In *Digital Scholarly Editing*, edited by Matthew James Driscoll and Elena

Pierazzo, 1st ed., 4:119-34. Theories and Practices. Open Book Publishers. <https://www.jstor.org/stable/j.ctt1fzhh6v.11>.

Turska, Magdalena, James Cummings, and Sebastian Rahtz. 2016. "Challenging the Myth of Presentation in Digital Editions." *Journal of the Text Encoding Initiative*, no. Issue 9 (September). <https://doi.org/10.4000/jtei.1453>.

Van Mierlo, Wim. 2022. "The Scholarly Edition as Digital Experience: Reading, Editing, Curating." *Textual Cultures* 15 (1): 117-25.

From dialect features to structured data

Modelling spoken Arabic varieties in the WIBARAB project

Veronika Engler (1);
Ana Iriarte Díez (2);
Karlheinz Mörrth (1);
Stephan Procházka (2);
Laurent Romary (3);
Daniel Schopper (1);
Kinga Sramó (1);
Katharina Wünsche (1)

(1) Austrian Academy of
Sciences, Austria;
(2) University of Vienna,
Austria;
(3) National Institute for
Research in Digital Science
and Technology, France

Keywords

linguistics, oriental studies, language
varieties, data modelling, visualisation

Abstract

The *Vienna Corpus of Arabic Varieties (VICAV)* is a language documentation platform that provides access to a growing collection of digital language resources. Integrating approaches from language technology and the wider field of text-oriented digital humanities, the project aims to address issues of representing heterogeneous data by providing a flexible, yet sustainable technical environment based on a modular data architecture. In addition to a bibliography of research literature, typologically similar to what can be found in the *World Atlas of Language Structures (WALS)* or the *Database of Arabic Dialects (DAD)*, VICAV offers several types of data: the so far largest part of the

collection comprises linguistic profiles (i.e. standardised concise descriptions of linguistic varieties), structured lists of linguistic features, sample texts, corpora of unmonitored speech, and dictionaries. The VICAV infrastructure is meant to ensure consistent encoding across projects as well as sustainable creation and publishing workflows across projects and builds largely on TEI (P5) as its underlying data model (Procházka et al. 2015).

The newest and, in terms of resources, duration, and scope, most extensive addition to the VICAV projects is the ERC Advanced Grant WIBARAB (What is Bedouin-type Arabic? 101020127-WIBARAB). It investigates the linguistic and socio-historical realities behind the millennia-old dichotomous distinction between Bedouin or sedentary dialects (Procházka 2024). The central component of WIBARAB is a linguistic feature database covering over 300 varieties of spoken Arabic. It incorporates data from hitherto little researched areas collected in campaigns in Saudi Arabia, Kuwait, Jordan, Lebanon, Sudan, Tunisia and Morocco, along with previously published material. The project has a strong focus on open access, data structures, standards and best practices, as well as data modelling.

By contrast to comparable other projects collecting linguistic data, WIBARAB has adopted a strictly text-oriented approach that has been grounded in the application of the *Guidelines of the Text Encoding Initiative (TEI)*. While TEI (P5) provides a well-tried inventory of elements to represent morphological and lexical concepts, the issue of representing other grammatical phenomena has received less attention and modelling extra-linguistic and socio-linguistic phenomena constitutes the interesting part of this encoding challenge. Our paper will focus on the WIBARAB team's customisation, trying to re-use as many concepts existing in the TEI guidelines as possible, document workflow steps such as the writing of a meaningful, re-usable ODD, and examine some encoding decisions by furnishing examples of phonological, morphological, syntactical, phraseological and lexical data.

Finally, we will touch on some methodological challenges in developing an efficient research-tool on the basis of this TEI-encoded, intricately structured linguistic database. It has already been used for a wide range of varied research questions such as the in-detail description of particular linguistic varieties (Torzullo 2024) as well as socio-linguistic approaches dealing with questions of intergenerational variation, intra-speaker variation and identity (Iriarte Díez 2025a, 2025b). We will also provide a first glimpse of the evolving map-based front-end which will be integrated into the overall VICAV infrastructure.

References

- Iriarte Díez, A. (2025a) ““Bedouins” in Beirut (and surroundings): Arab Khalde – Intergenerational linguistic variation in a traditionally seminomadic community 12 km south of Beirut”, in A. Iriarte Díez and St. Procházka (eds.) *What is Bedouin-Type Arabic? New Data and Fresh Perspectives. Wiener Zeitschrift für die Kunde des Morgenlandes (WZKM)* 115.
- Iriarte Díez, A. (2025b) “Intra-speaker variation and identity performance: Two texts in the Arabic of Arab Khalde (Lebanon)”, in A. Iriarte Díez and St. Procházka (eds.) *What is Bedouin-Type Arabic? New Data and Fresh Perspectives. Wiener Zeitschrift für die Kunde des Morgenlandes (WZKM)* 115.
- Procházka, S. (2024) “How solid is the linguistic basis for the Bedouin sedentary split used in the classification of Arabic dialects? ”, in C. B. Ramos, J. Guerrero and M. B. Fernández (eds.) *AIDA Granada: A pomegranate of Arabic varieties. Zaragoza: Prensas de la Universidad de Zaragoza: 359-370. <https://phaidra.univie.ac.at/o:2108301>*.
- Procházka, S. and K. Moerth (2015) “The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects”, in M. Al-Hamad, R. Ahmed and H. Aloui (eds.) *Lisan*

Al-Arab: Studies in Contemporary Arabic Dialects, Proceedings of the 10th International Conference of AIDA, Qatar University 2013. Vienna: LIT Verlag: 209-218.

Torzullo, A. (2024) “The Arabic Dialect of a Jordanian Camel-Breeder Tribe: a Comparative Analysis of Selected Phonological and Morphological Features of the Bani Šaxar Variety”, in C. B. Ramos, J. Guerrero and M. B. Fernández (eds.) *AIDA Granada: A pomegranate of Arabic varieties. Zaragoza: Prensas de la Universidad de Zaragoza: 409-420. <https://phaidra.univie.ac.at/o:2108301>*.

From Practice to Framework

Model Digital Scholarly Editions in the Jagiellonian Digital Platform

Magdalena Eulalia
Komorowska;
Iwona Grabska-Gradzińska;
Joanna Hałaczekiewicz

Jagiellonian University,
Poland

Keywords

Digital Scholarly Editions, TEI XML, Digital Humanities Infrastructure, Jagiellonian Digital Platform

Abstract

Entering New Territories of Digital Scholarly Editing

The Digital Scholarly Editing Laboratory (LabEdyt), an interdisciplinary team within the flagship project Digital Humanities Lab at the Jagiellonian University, was founded in 2022 to address the growing need for sustainable digital scholarly editions (DSEs) in Poland. While DSEs are already being developed with some regularity across many academic contexts, they remain a complex and evolving challenge – even in well-established environments. In the Polish context, their adoption is still limited to a few research centers and often regarded as too complicated to be worth considering. Persistent myths and misconceptions about the usefulness and durability of such editions, along with a reluctance to publish under open licenses, still act as a barrier to progress. *LabEdyt's mission is to create practical workflows*

and reusable models that lower the entry barrier to digital editing and support researchers in transitioning from print-based to digital methods. This paper presents the methodological framework and current outcomes of three model projects developed within the Jagiellonian Digital Platform.

Mapping Workflows and Good Practices in Digital Editing

The first project involves the revitalization of “Neolatina Sarmatica”, a digital collection of Latin texts by authors associated with the Polish-Lithuanian Commonwealth. Originally created between 2006 and 2011, the collection had become obsolete and lacked structured markup. By 2022, the existing website was no longer practical to use, as advances in web technologies had rendered it outdated. Such obsolescence of research websites – including digital scholarly editions – is a widespread problem. The pilot project on “Neolatina” therefore aimed not only to make this collection accessible to readers again, but also to develop guidelines for creating editions that remain sustainable over time and for reviving dormant digital projects. Using web scraping and TEI XML conversion, the team rebuilt the edition in TEI Publisher integrated with an external database. The new platform now supports side-by-side views of transcription, facsimile, and translation, along with interactive annotations. This project demonstrates the long-term value of adhering to encoding standards and underscores the importance of sustainable design for digital editions.

The second project focuses on “Moralia”, a 17th-century manuscript by Wacław Potocki. The work of this Baroque poet comprises more than 2,100 pieces and was modeled on Erasmus of Rotterdam’s *Adagia*. Its considerable scale has long posed a challenge for editors, and it has never received a complete scholarly edition. The necessity of working with the manuscript also provided an opportunity to test a workflow incorporating a handwriting text recognition (HTR) tool. The team trained a custom handwriting recognition model in Transkribus on more

than 100 annotated pages, achieving over 97% transcription accuracy. Current work involves developing a large language model capable of producing modernized versions of the text in line with the editorial principles. The model will be designed not only to 'translate' the diplomatic transcription into a modernized text, but also to automatically encode these changes in TEI. The edition will also integrate intertextual references, most notably a surviving copy of Erasmus's *Adagia* annotated by Potocki himself. These materials will be re-framed in a digital commentary format through the use of structured markup and hypertext techniques.

The third project aims to develop a method for constructing digital editions from materials originally prepared for print. As a case study, the team is working on a scholarly edition of 20th-century Szymon Laks's correspondence with Krystyna and Czesław Bednarczyk, which had been typeset in Adobe InDesign and published in 2018. The project explores workflows for mapping style-based formatting onto XML structures. While certain structural elements can be retroconverted automatically, others – such as personal names or metadata – require manual tagging, custom scripting, or the use of AI-assisted tools like Named Entity Recognition. The challenges we continue to encounter reveal broader issues in digital retroconversion, particularly when editorial markup was not anticipated during the print-production stage.

Building a Community of Practice

All three projects are being released incrementally via the Jagiellonian Digital Platform (<https://labedyt.dhlab.uj.edu.pl>), which serves both as a stable public repository and as an internal development space for digital editors at the Jagiellonian University. Rather than inventing new proprietary tools, LabEdyt adapts and extends open-source solutions such as TEI Publisher × Jinks, contributing to their international development through applied use and structured feedback.

In line with the “New Territories” theme, LabEdyt's work represents an effort to establish digital scholarly editing as a viable and sustainable practice in Poland. By documenting challenges, sharing workflows, and publishing reusable models, the team supports a growing community of researchers who are ready to explore new directions at the intersection of textual scholarship and digital infrastructure.

Functors and format conversion

Riccardo Del Gratta;
Angelo Mario Del Grosso

Institute for Computation
Linguistics A. Zampolli-
CNR-ILC, Italy

Keywords

Interoperability, Category Theory, Text
Analysis, Computational Linguistics,
Linked Open Data

Abstract

This contribution highlights the potential of functorial format conversion in applications such as digital philology and structured text analysis, where the heterogeneity of document formats often interferes with interoperability among text analysis tools.

The authors introduce a formal framework based on Category Theory, using *functors* as an abstract means for managing format conversion. Functors maintain coherence during transformation, enabling tool pipelining even when format mismatches occur.

In the proposed framework, functors enable the transformation of documents and their associated processing tools from one category (containing documents in format *f*) into another (in format *g*), preserving fundamental structural and operational properties. The idea is to represent format conversion not as a technical task but as a structured mapping that maintains the algebraic properties of the source and target domains.

In previous works, the authors illustrated the practical implications of this approach through working examples that convert

annotated XML into JSON documents, enriching them with externally linked data.

This contribution underscores that beyond format conversion, the framework shows further research questions such as (i) document isomorphism; (ii) partial structure mapping, and (iii) natural transformations between functors.

(i) and (ii) address the following scenario: First, an XML document *DO_x* is “functored” to a JSON document *DO_j* using functor *F*. Then, *DO_j* is “functored” into a possibly new XML document *D1_x* with functor *G*. In (i), we investigate suitable metrics to evaluate the similarity between *DO_x* and *D1_x*; in (ii), we study the implications of using functors that “forget” part of the original structure, i.e., the directive about the XML schema declaration. Finally, in (iii), we analyze the application of natural transformations to manage the mapping of *DO_x* into two JSON documents with different structures. (iii) is relevant to handle the “diffing” issues on documents that serialize the same content with different structures.

References

- R. Del Gratta, F. Boschetti, L. Bambaci, and F. Sarnari, “Document analysis and Textual philology: A Formal Perspective,” *International Journal of Information Science and Technology*, vol. 5, no. 1, pp. 5–15, 2021. [Online]. Available: <https://www.innove.org/ijist/index.php/ijist/article/view/192>
- , “Approaching document analysis with a formal model,” in 6th International IEEE Colloquium on Information Science and Technology, Agadir, Morocco, 2020, pp. 208–214.
- R. Del Gratta, F. Boschetti, A. Del Grosso, S. Zenzaro, and L. Bambaci, “Philology as a dynamic system,” *Umanistica Digitale*, vol. 6, p. 1–20, Jan. 2022. [Online]. Available: <https://umanistica-digitale.unibo.it/article/view/13684>

- R. D. Gratta, S. Zenzaro, A. D. Grosso, and F. Boschetti, "Category theory, Document Analysis, and Philological Operations. A formal approach: limitations, and challenges," *International Journal of Information Science and Technology*, vol. 9, no. 1, pp. 11-20, 2025. [Online]. Available: <https://innove.org/ijist/index.php/ijist/article/view/288>
- S. Awodey, *Category Theory*, 2nd ed. New York, NY, USA: Oxford University Press, Inc., 2010.
- S. Mac Lane, *Categories for the Working Mathematician*, 2nd ed., ser. Graduate Texts in Mathematics. Springer, 1998. [Online]. Available: <http://www.worldcat.org/isbn/0387984038>
- E. Riehl, *Category Theory in Context*, ser. Aurora: Dover Modern Math Originals. Dover Publications, 2017. [Online]. Available: <https://books.google.it/books?id=6B9MDgAAQBAJ>
- J. Gerrans and R. S. Sherratt, "Comparing xml and json characteristics as formats for data serialization within ultralow power embedded systems," *IEEE Embed. Syst. Lett.*, vol. 16, no. 4, p. 489-492, Dec. 2024. [Online]. Available: <https://doi.org/10.1109/LES.2024.3450576>
- S. Zunke and V. D'Souza, "JSON vs XML: A comparative performance analysis of data exchange formats," *IJCSN International Journal of Computer Science and Network*, vol. 3, no. 4, pp. 257-261, 2014.
- P. C. Lalith, S. Goel, M. Kakkar, and S. Sharma, "Simplifying code translation: Custom syntax language to c language transpiler," in *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 2025, pp. 1-6.
- A. Okutan, S. Merten, C. C. Michael, and B. Ryjikov, "Leveraging rag-llm to translate c++ to rust," in *2024 International Conference on Assured Autonomy (ICAA)*, 2024, pp. 102-105.

Generating TEI Documents Through a Game of Dice

Viktor J. Illmer; Nele Heindorf;
Roya Zendeбудіe; Mark Schwindt;
Frank Fischer

Freie Universität
Berlin, Germany

Keywords

German Literature, Generative Literature,
Drama, Gamification, 19th Century

Abstract

In 1829, German author Georg Nikolaus Bärmann published a 272-page book entitled "Neunhundert neun und neunzig und noch etliche Almanachs-Lustspiele durch den Würfel" ("Rolling the Dice for 999 and Many More Almanac Comedies"). The work consists of 1,200 randomly printed text pieces, mostly dialogs between the five main characters (the uncle, the niece, the maid, the lover, his servant). With the help of a dice, you can put together an entire German-language one-act comedy from these snippets. A table is provided for this purpose, in which a piece of text is assigned to each roll of a total of 200 rolls, from which a coherent play is then created. Mathematically, this results in 6 to the power of 200 possibilities and therefore more plays than atoms in the universe (Baker 2021).

We first digitized Bärmann's book using the "German Fraktur 18th Century - WrDiarium_M9" model on Transkribus and corrected it on Wikisource according to the four-eyes principle. We then produced a TEI version of the book featuring all 1,200 text snippets. On this basis, the book was turned into a web

app and the game can now be played online (Illmer et al. 2025). It is available at <https://temporal-communities.github.io/999/>.

For the TEI2025 conference, we developed an extension to the app that allows you to download a randomly diced game in TEI format. This enables extensive analysis of the randomly generated plays. We present the TEI component of the game app as well as quantitative and stylometric experiments that we conducted with the generated plays.

To facilitate the understanding of the play - written in a highly idiosyncratic German with historical orthography - we also provide an AI-assisted translation to English of the entire play. The translation is also encoded in TEI and can be accessed from within the app.

To sum up, with this paper we will demonstrate how TEI can be used to encode a literary game and serve as a basis for an interactive web app, and we will discuss the implications for the TEI-encoding of AI-assisted translations.

References

- Harry Baker: How many atoms are in the observable universe? In: Live Science, July 10, 2021. Available: <https://www.livescience.com/how-many-atoms-in-universe.html>
- Georg Nicolaus Bärmann: Neunhundert neun und neunzig und noch etliche Almanachs-Lustspiele durch den Würfel. Zwickau: Schumann 1829. (digitized version: https://de.wikisource.org/wiki/Neunhundert_neun_und_neunzig_und_noch_etliche_Almanachs-Lustspiele_durch_den_W%C3%BCrfel)
- Viktor J. Illmer, Frank Fischer, Mark Schwindt, Jonas Rohe: »999 und noch etliche [mehr]«. Georg Nikolaus Bärmanns »Würfel-Almanach« von 1829 als Web-App. In: DHd2025: »Under Construction«, 3-7 March 2025. Book of Abstracts. Bielefeld University. <https://doi.org/10.5281/zenodo.14943242>

Generative AI for XML-TEI Encoding

Exploring the potential of LLMs for Spanish Golden Age Theatre

Marco De Cristofaro (2,4);
Emanuele Leboffe (1);
Daniel Zilio (3)

(1) Universitat Autònoma de Barcelona, Spain;
(2) Université de Mons;
(3) Università degli studi di Padova;
(4) Université de Namur

Keywords

Spanish Golden Age theater, digital critical editions, Generative Artificial Intelligence, Large Language Models, Digital Humanities

Abstract

Nowadays, encoding texts in XML-TEI has become a fundamental practice in the field of digital humanities. Its importance lies in creating structured, interoperable representations of texts, enabling their storage in online databases where they can be accessed, consulted, and reused by other scholars, for example, to conduct large-scale quantitative analyses such as automated data extraction or linguistic profiling.

However, encoding can be particularly challenging, especially when the texts involved display significant complexity. Spanish Golden Age comedies offer a fitting example. These texts present a layered structure composed of multiple elements – characters' interventions, verses, stanzas, stage directions – all of

which require detailed markup. If we also take into account the considerable length of these plays – each averaging around 3,000 verses – encoding can quickly become an arduous and highly mechanical task, threatening the sustainability of research unless effective methods for automating or facilitating the workflow are adopted.

Given these complexities, generative artificial intelligence stands out as a promising solution. Its potential for academic research has already been demonstrated in the sciences (Schmidgall et al. 2025) and the humanities (DeRose 2024); moreover, recent developments in AI have influenced how archival materials and literary texts are preserved and accessed (Colavizza et al. 2021), highlighting its effectiveness in supporting the study and conservation of literary archives and cultural heritage more broadly (Carbé 2023).

With a view to ensuring research sustainability, this paper proposes an approach to automatic XML-TEI encoding of theatrical texts, using as a case study the approximately 350 Spanish Golden Age plays already encoded and made available by the Biblioteca Digital ARTELOPE²³.

In the initial phase, we preprocessed the material in two parallel ways: first, we generated simplified XML-TEI files by stripping away non-essential TEI elements; second, we produced plain TXT files containing the same textual content. This dual preparation served different purposes: one portion of the original XML files was used to provide structured references for the model during training and prompt design, while another portion was held back and used as benchmark data to evaluate the accuracy of the final automatic encoding process. To produce these automatic encodings, the plain text files were fed into two different large language models – OpenAI ChatGPT-4 and

Anthropic Claude 3.5 Sonnet – using both chat-based interaction and API-based workflows.

The chat-based approach revealed a relative independence between the two models, with a very high recognition rate, although consistency tended to decline over the course of the text. The API-based approach, on the other hand, showed greater adherence to the intended encoding. While in both cases we relied on manual verification, for long-term research sustainability, developing a model for automatic validation based on standardized metrics will be crucial.

In conclusion, the first results suggest that LLMs hold considerable promise for the automatic encoding of texts characterized by significant length and complexity. Although further refinement is needed, these indications are encouraging and point to generative AI as a potential breakthrough for large-scale philological projects.

23 The corpus of comedies encoded in XML-TEI is available at: https://gitlab.com/artelopez/ARTELOPE/-/tree/main/XML-TEI%20Play-texts?ref_type=heads (last accessed: 10/07/2025).

Graph-Based Digital Editions and TEI

Towards Interoperable and Assertive Scholarly Editing

Sebastian Enns;
Stefan Armbruster;
Andreas Kuczera

TH Mittelhessen, University of
Applied Sciences, Germany

Keywords

Graph-based Edition, TEI/XML, ATAG,
Annotation, Assertive Editing

Abstract

The concept of the *assertive edition*, as introduced by Vogeler (2019), places the semantic interpretation of textual content at the center of scholarly editing. It entails not only the transcription of historical sources but also the representation of interpretive statements within the data structures of a digital edition. While the TEI guidelines offer means for annotating named entities and events, they are typically implemented in XML, which follows the *Ordered Hierarchy of Content Objects (OHCO)* model. This structure makes it difficult to express overlapping hierarchies, stand-off annotations, or semantically nuanced assertions without relying on complex and often project-specific workarounds (cf. Vogeler 2021; cf. Stoyanova 2023; cf. Kuczera 2022; cf. Cugliana et al. 2024b).

In response to these structural limitations, graph-based approaches have been developed. Kuczera (2016, 2024)

introduced *Applied Text As Graph (ATAG)*, a model in which the text is structured as a linear sequence of character-level nodes. Annotations are implemented as dedicated nodes that span sections of text and may include their own content (cf. Kuczera/Neill 2019; cf. Kuczera 2024). This allows for overlapping annotation layers and recursive structures while maintaining a distinction between transcription and interpretation. In contrast to XML-based models, ATAG separates textual sequence from editorial semantics at the structural level (cf. Kuczera 2024).

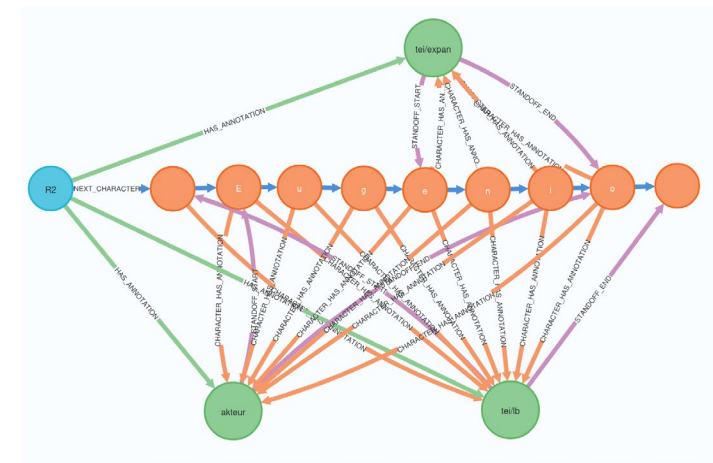


Fig. 1: A labeled-property graph visualizing the text “Eugenio” with letters as orange nodes connected to green annotation nodes, which are linked to a text node (Kuczera 2024).

Building on these developments, a domain-independent metamodel for digital editions is currently being developed in Sebastian Enns’ ongoing doctoral research titled *Management of Heterogeneous Data in Digital Editions: A Model-Driven Approach*. The metamodel supports the construction of domain-specific models and promotes semantic consistency

across diverse editing contexts. The conceptual scope of this approach is further illustrated in Cugliana et al. (2024b), where graph-based abstractions are used to formalize coexisting hierarchies, semantic layers, and variant readings. While the contribution remains theoretical, it highlights the potential of graph technologies to address representational limitations in XML-based models.

This paper focuses on the import of existing TEI/XML data into a graph-structured TEI environment based on ATAG. In the digital edition *The Socinian Correspondence*, transcriptions created in the Oxygen XML editor are processed using a TEI/XML parser developed for this purpose. The parser interprets the encoded TEI structures and translates them into a graph representation that preserves both sequence and annotation logic. Structural and interpretive elements from the TEI/XML input are mapped to graph components aligned with ATAG and organized according to a domain-specific model derived from the metamodel. The paper outlines this import pipeline, the mapping strategies involved, and the resulting publication setup based on graph-structured data.

This approach situates TEI within a broader editorial context. By abstracting TEI-encoded content into graph-based representations, it becomes possible to realize assertive, algorithmic, and interoperable digital editions that open up *new territories* for modeling textual structure and editorial interpretation.

References

Cugliana, Elisa / Enns, Sebastian / Kuczera, Andreas (2024a): "Sortes Dictae Sunt. Methods for Editing Mediaeval Books

of Fortune." In: *Zeitschrift für digitale Geisteswissenschaften* 9 (2024). DOI: https://doi.org/10.17175/2024_005.

Cugliana, Elisa / Ward, Aengus / van Zundert, Joris J. / Kuczera, Andreas / Grüntgens, Max (2024b): "Computational Approaches and the Epistemology of Scholarly Editing." *International Journal of Digital Humanities*. DOI: <https://doi.org/10.1007/s42803-024-00088-z>.

Kuczera, Andreas (2016): "Digital Editions beyond - XML-Graph-based Digital Editions. In: Proceedings of the 3rd Histoinformatics Workshop on Computational History. (Histoinformatics: 3, Krakow, 07.11.2016) Aachen, 2016. Published at https://ceur-ws.org/Vol-1632/paper_5.pdf.

Kuczera, Andreas (2022): "TEI Beyond XML - Digital Scholarly Editions as Provenance Knowledge Graphs." In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): *Graph Technologies in the Humanities - Proceedings 2020*, published at <http://ceur-ws.org/Vol-3110>.

Kuczera, Andreas (2024): "Applied Text as Graph (ATAG)". DHd 2024 Quo Vadis, Passau, Deutschland. Zenodo. DOI: <https://doi.org/10.5281/zenodo.10698323>.

Kuczera, Andreas / Neill, Iian (2019): "The Codex - An Atlas of Relations." In: Kuczera, A., Wübbena, T., Kollatz, T., editors, *Die Modellierung des Zweifels - Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, Volume 4, Special Issue of *Zeitschrift für digitale Geisteswissenschaften*. DOI: https://doi.org/10.17175/sb004_008.

Stoyanova, Silvia (2023): "Articulating Intra- and Intertextual Relationships in the Fragment Collection. Working with the Digital Edition of Giacomo Leopardi's Zibaldone." *magazén* 4(1), p. 13-42. DOI: <http://doi.org/10.30687/mag/2724-3923/2023/07/001>.

Vogeler, Georg (2019): "The 'assertive edition' On the consequences of digital methods in scholarly editing for historians."

International Journal of Digital Humanities 1 (2019), p. 309-322.

DOI: <https://doi.org/10.1007/s42803-019-00025-5>.

Vogeler, Georg (2021): "Standing-off Trees and Graphs: On the Affordance of Technologies for the Assertive Edition."

In: *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, herausgegeben von Elena Spadini, Francesca Tomasi und Georg Vogeler, p. 73-94. Norderstedt: Books on Demand (BoD). ISBN: 978-3-7543-4369-2.

Haute couture (for the masses)

Magdalena Turska

e-editiones, Poland

Keywords

TEI Publisher, Jinks, digital scholarly editions

Abstract

Pierazzo was not the first one to point out that digital editions typically require high-tech skillset and a big upfront investment, not to mention the compound cost of maintenance. She also coined the opposition of *haute couture* vs *prêt-à-porter* editions: the latter would allow for streamlined production and publication of editions attainable for anyone without significant funding and institutional support – at the cost of dealing with necessary limitations imposed by the generic publication platforms. Numerous voices of the textual scholars criticized this approach, among them Cunningham, who deemed that “*to wish for a one-size-fits-all software for the production of DSE’s is and would be misguided*”. Is there a way to accommodate the needs and wishes of editors at the same time taking into account the technical and economic sustainability as well as future re-use and interoperability of the DSE’s data?

I would like to present lessons learned as well as sustainability and interoperability implications from more than two dozen heterogeneous editorial projects, all realized according to the design principles heavily influenced by literate programming

in general and the TEI Processing model in particular. I will try to prove that despite the great diversity of source material, research questions and scholarly domains, combined with similarly varying XML encoding flavours, a highly customized *haute couture* editions can be created within the same framework. Analysis of the internal structure of application and data packages will demonstrate overwhelming overlap in implementation even between projects as remote as genetic edition of Andersen's fairy tales, register of Older Slovenian Manuscripta, correspondence of van Gogh and the Lexicon of the Greek Personal Names.

References

- Pierazzo Elena, *What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter*, „International Journal of Digital Humanities” 1 (2) 2019.
- Cunningham Richard, *Theorizing a Digital Scholarly Edition of Paradise Lost*, *Advances in Digital Scholarly Editing: Papers Presented at the Dixit Conferences in the Hague, Cologne, and Antwerp*, 2017.

Introducing UDraCor

The Ukrainian Drama Corpus in TEI

Frank Fischer (1);
Julia Havrylash (2);
Daniil Skorinkin (3);
Mark Schwindt (1)

(1) The Free University of Berlin,
Germany;
(2) University of Trier, Germany;
(3) University of Potsdam,
Germany

Keywords

drama, drama encoding, dracor, Ukrainian
language, Ukrainian literature

Abstract

The DraCor platform (dracor.org), grounded in the concept of “programmable corpora” (Fischer et al. 2019), offers a robust infrastructure for hosting, accessing, and analysing TEI-encoded dramatic texts. With over 4,330 plays across 21 corpora in 17 languages on its production instance (and many more in the making), DraCor has evolved into an open, research-oriented environment for both humanistic inquiry and computational analysis. While early corpora were initiated by the core development team, more recent additions – including Alsatian, Czech, Dutch, Hungarian, Polish, and Yiddish drama – reflect the increasing role of community-driven contributions and cross-linguistic collaboration.

In response to Russia's full-scale invasion of Ukraine in 2022 – a war waged not only against lives but also against cultural identity – the DraCor team launched the Ukrainian Drama Corpus (UDraCor). This initiative exemplifies the expansion of

TEI-based infrastructure into previously underrepresented regions and demonstrates the resilience of collaborative encoding efforts under challenging circumstances. The corpus is curated by a network of Ukrainian and international scholars and includes 19th- and early 20th-century Ukrainian plays drawn from a range of digitized but largely unstructured sources.

UDraCor also serves as a testing ground for innovative encoding practices. The development of EzDrama, a lightweight TEI/XML conversion script (Skorinkin 2024), and the integration of LLM-assisted encoding pipelines (Skorinkin & Giovannini 2024) have enabled a scalable, semi-automated workflow for transforming raw text into TEI/XML. EzDrama was originally created for UDraCor, but since then helped a number of scholars and volunteers encode plays in a variety of languages, including Czech, German, and Yiddish. Owing to its good alignment with the original raw text, EzDrama serves as an intermediary format for LLM-based encoding, helping control for hallucinations and omissions. These methods lower the entry threshold for corpus contributors and accelerate the growth of TEI corpora.

We illustrate UDraCor's research potential with a case study of Mykola Kulish's play 97—originally titled *Hunger*, and set during the 1921–22 famine in the Kherson region. Forced to alter the title and the play's ending under political pressure, Kulish nonetheless encoded a stark socioeconomic conflict into the character list itself, explicitly distinguishing between rich and poor villagers. This authorial metadata offers a rare “ground truth” for validating computational models of social network clustering in dramatic texts.

UDraCor thus embodies the central themes of TEI 2025: it extends the TEI community into new geopolitical and scholarly territories, showcases sustainable and community-based corpus development, and demonstrates the integration of novel, scalable encoding methodologies. Most importantly, it contributes to the preservation and international visibility of Ukrainian literary heritage at a moment when such work carries both scholarly and cultural urgency.

References

- Fischer, F. *et al.* (2019) ‘Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama’. *Digital Humanities 2019: ‘Complexities’ (DH2019)*, Utrecht: Zenodo. Available at: <https://doi.org/10.5281/zenodo.4284002>.
- Skorinkin, D. (2024) ‘EasyDrama: a lightweight solution for encoding plays in TEI-XML’, in *Book of Abstracts of TEI 2024. TEI 2024*, pp. 38–39. Available at: <https://zenodo.org/records/13883242>.
- Skorinkin, D. and Giovannini, L. (2024) ‘Towards an LLM-powered encoding workflow for plays’, in *Book of Abstracts of TEI 2024. TEI 2024*, pp. 94–95. Available at: <https://zenodo.org/records/13883242>.

Ioannes Dantiscus' Itinerary Online

Anna Skolimowska;
Katarzyna Jasińska-Zdun;
Mateusz Materek

University of Warsaw,
Poland

Keywords

interactive itinerary, interactive map,
Renaissance, Dantiscus, TEI Publisher

Abstract

Launched in May 2025, the *Ioannes Dantiscus' Itinerary* offers an interactive digital resource illuminating the travels of the Renaissance diplomat, humanist, and poet. Built upon the *Corpus of Ioannes Dantiscus' Texts and Correspondence* (CIDTC), the Itinerary draws from a MySQL database, and XML files encoded in TEI.

Each Itinerary “entry” is defined by a specific date and place, and includes information on relevant sources, often linked to digital editions. These entries are also visualized through the *Ioannes Dantiscus Itinerary Web Map*, which displays all documented locations of Dantiscus's presence, along with selected, verified routes. Users can navigate both the Itinerary and the Map using filters and adjustable date ranges.

The Itinerary website was developed in eXist and TEI Publisher, a platform dedicated to digital editions, while the Map uses Node.js and QGIS. For interoperability—particularly with the Itinerary Web Map and other CIDTC components—data is exported from MySQL in JSON format. Within the eXist framework, JSON is transformed into TEI XML using XQuery.

The TEI Publisher serves here not as a platform for displaying primary texts, but as a metadata aggregator and visualisation interface. It links entries defined by time and place to corresponding CIDTC records (e.g., documents, text fragments), enabling analyses and insights that would be difficult to get reading the corpus alone.

The focus is on establishing connections across datasets, positioning the Itinerary as an important node in a larger digital ecosystem. Given the outdated infrastructure of CIDTC, the use of TEI Publisher is also seen as an exploratory step toward potentially migrating the project's full dataset to an eXist-db-based environment.

Jinks / TEI Publisher 10

Magdalena Turska

e-editiones, Poland

Keywords

TEI Publisher, editing workflow, digital editions

To acknowledge growing community of TEI and Publisher users in Poland, we'd like to offer this workshop primarily in Polish, with "whispered translation" offered for non-Polish speakers. In case of high interest, we could also cover two separate language tracks, Polish and English.

Abstract

Due to our modular approach to crafting software, TEI Publisher grew from just two core packages to an entire framework of modules, libraries and repositories, independently released with their own version numbers. Nevertheless, community still likes to think of major new versions of the TEI Publisher. The 10th, anniversary edition, marks a decade since the first Publisher presentation (and our first workshop) at the TEI conference in Lyon.

To celebrate the occasion, we would like to present, in the hands-on format, the current state of TEI Publisher's development, in particular the new Jinks application manager. In practice, Jinks is an edition builder where user can click together an edition and tailor it to her needs in literally minutes, thanks to a new Jinks library of profiles and features. Existing features already cover basic needs of every edition plus quite some very specific requirements, for e.g. parallel views, correspondence or presentation of textual variants. As usual, everyone is invited to bring forward other features to be added to this community-curated collection of building blocks for scholarly publications.

JinnTap

A browser-based TEI-XML editor

Wolfgang Meier (1,2,4);
Lars Windauer (1,2);
Joseph Wicentowski (3)

(1) e-editiones, Switzerland;
(2) JinnTec GmbH, Germany;
(3) FSI Office of the Historian,
USA;
(4) Heidelberger Akademie der
Wissenschaften, Germany

Keywords

TEI, editor, workflow

Abstract

The formalisation of an edition in an informative structure, e.g. via a mark-up language, that enables its processing is, for many scholars, one of the most challenging aspects of digital editing. This is why several endeavours exist that facilitate textual encoding (such as ediarum²⁴, eLaborate²⁵, FairCopy²⁶, TEI Publisher²⁷ or TextGrid²⁸ to name a few). They aim at assisting humanists in different parts of the editorial pipeline, including transcription, semantic annotation and other aspects.

²⁴ <https://www.bbaw.de/bbaw-digital/telota/forschungsprojekte-und-software/ediarum>

²⁵ <https://www.elaborate.huygens.knaw.nl/>

²⁶ <https://faircopyeditor.com/>

²⁷ <https://teipublisher.org>

²⁸ <https://textgrid.de/>

In this demo we will show JinnTap²⁹, a means to edit TEI-XML documents using a browser-based rich text editor, and highlight the differences to existing initiatives. While JinnTap was initially inspired by concrete workflow needs at the Office of the Historian, it is intentionally designed to be broadly applicable.

JinnTap preserves the structure of the XML in the editor, thus rooted in the WYSIWYM paradigm, but exploiting the common features of WYSIWYG editors. JinnTap forces the editor to critically reflect on the data model, and create a coherent customization of the TEI schema, tightly fitting the project. While still under development, the default schema already supports:

- block level elements like headings, paragraphs, lists, divisions
- elements for arbitrarily nested inline formatting like `<hi>` or `<title>`
- semantic markup for people, organizations, places, terms
- analytic elements like `<date>`
- footnotes
- inline nodes representing an alternative like `<choice>`, `<abbr>`, `<expan>`
- figures, figure descriptions and graphics

The schema is fully customizable as long as the distinction between block and inline elements is clear, which we argue is an asset in terms of modelling and not a limitation.

²⁹ <https://github.com/JinnElements/jinn-tap>

LEGOstyle

Building modular, flexible Editions with TEI

Thomas Kollatz

Academy of Sciences and Literature
Mainz, Germany

Keywords

Digital Edition, Correspondences, CMIF

Abstract

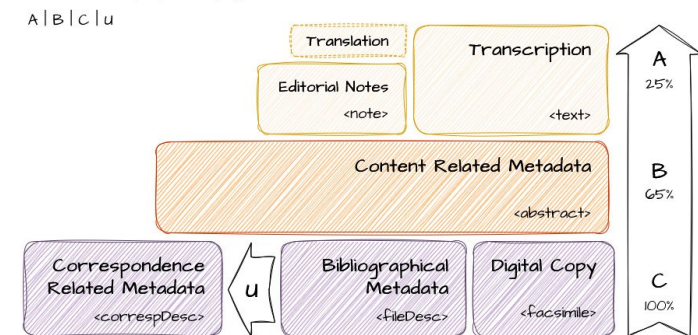
The 'Buber-Correspondences Digital' (BKD) project focuses on the 43,000 preserved letters, postcards and telegrams exchanged between the philosopher of religion, Martin Buber (1878-1965), and his over 7,000 correspondents. Based in Mainz and Frankfurt am Main, this long-term project is funded by the Academies' Program of the Union of German Academies (AGATE 2025). During the editing process, all correspondence is annotated with the TEI. However, providing a full-text edition is beyond the scope of this project due to the collection's vast size. Nevertheless, the concept of a modular edition with varying levels of granularity is a viable option.

The modular edition

BKD has introduced a modular 'building block system' into the editorial process. The building blocks, referred to as 'A', 'B', 'C' and 'U', respectively represent varying levels of indexing (BKD 2025: 2.2. Erschließungskategorien - Die modulare Edition). The basic 'C' block is included in every edited source (100% of

correspondence). It contains fundamental information on the sender, recipient and date, i.e. the TEI element `<correspDesc>`, as well as information on the extent, object type, mode of writing, holding institution and signature. This enables all data to be entered into the web service correspSearch ('correspSearch - Briefeditionen vernetzen (3.1.0)' 2025) and to be linked with other correspondence editions. The 'B' building block serves to introduce a dynamic element (Jurst-Görlach and Kollatz 2023). It contains a human-readable abstract (`<abstract>`) as well as lists of entities (persons, organizations, places, events and works) mentioned, and expresses existing relationships between these entities where applicable. The TEI element `<relation>`, which is used to generate triple-like statements, is employed for this purpose. This particular building block, which is also applied to 'A' letters, is used for 65% of all correspondence (Jurst-Görlach 2024). Building block 'A' is included in 25% of edited letters; contained in the TEI element `<text>`, there is a 'classic' full-text edition in which the correspondence is transcribed, annotated, and, if necessary, translated. Building block 'U' contains correspondence items that have not been preserved, but whose

LEVELS OF INDEXING



Graphic created by Denise Jurst-Görlach 2025 (CC-BY)

existence is proven by clear references in existing correspondence or other sources.

Conclusion

This LegoStyle approach enables flexibility. All information is provided in a standardized ODD-powered scheme, and upgrades can be made at any time by the project team or via reuse by third parties. For example, version 2 of the Correspondence Metadata Interchange Format ('CMIF V2' 2023) provides an ideal vocabulary for expressing the BKD approach (Dumont et al., no date): `cmif:Record` equals building block 'C', `cmif:Abstract` equals building block 'B', `cmif:Transcription` equals building block 'A', and `cmif:NoTextBase` equals building block 'U'.

The generated data is made available in the BKD repository from the outset ('Repositorium des Akademievorhabens Buber-Korrespondenzen Digital' 2022).

References

- AGATE. 2025. „Buber-Korrespondenzen Digital. Das Dialogische Prinzip in Martin Bubers Gelehrten- und Intellektuellennetzwerken im 20. Jahrhundert“. A European Gateway for the Academies of Sciences and Humanities. <https://agate.academy/id/PR768>.
- BKD. 2025. „Buber-Korrespondenzen Digital. Das Dialogische Prinzip in Martin Bubers Gelehrten- und Intellektuellennetzwerken im 20. Jahrhundert (BKD). Editionsrichtlinien und Schema“. https://adwmainz.pages.gitlab.rlp.net/digicademy/bkd/correspondences/schema/tei_bkd.html.
- CMIF V2. 2023. https://github.com/TEI-Correspondence-SIG/CMIF/blob/main/proposals/CMIFv2_proposal.md.
- correspSearch - Briefeditionen vernetzen (3.1.0). 2025. Edited by Stefan Dumont, Sascha Grabsch, Jonas

Müller-Laackman, Ruth Sander and Steven Sobkowski. Berlin-Brandenburgische Akademie der Wissenschaften 2025. <https://correspSearch.net>.

Dumont, Stefan, Ingo Börner, Dominik Leipold, Martin Anton Müller, Jonas Müller-Laackman, Klaus Rettinghaus, Gerlinde Schneider, Torsten Schrade, Peter Stadler, and Jakub Šimek. o. J. „DRAFT: Correspondence Metadata Interchange Format (CMIF) vocabulary“. TEI Correspondence SIG. <https://lod.academy/cmif/vocab/terms/>.

Jurst-Görlach, Denise. 2024. „Relationen dokumentieren in TEI - Kommentieren im Semantic Web“. In *KONDE Weißbuch*. Edited by Selina Galka and Helmut W. Klug in collaboration with Susanne Höfer in the project „Enlarging ‚Weißbuch Digitale Edition‘“. <https://www.digitale-edition.at/o:konde.250>.

Jurst-Görlach, Denise, and Thomas Kollatz. 2023. „Text Encoding without //text. The use of //abstract as means to avoid the one-dimensionality of ego-networks in ›Buber-Korrespondenzen Digital‹ project“. Paderborn. <https://gitlab.rlp.net/adwmainz/digicademy/bkd/bkd-presentations/teimec>.

Repositorium des Akademievorhabens Buber-Korrespondenzen Digital 2022. <https://gitlab.rlp.net/adwmainz/digicademy/bkd/correspondences>.

Linking Your *-ographies

Developing project-specific TEI Authority File Lookups for LEAF-Writer

James Cummings (1);
Luciano Frizzera (2);
Diane Jakacki (3);
Susan Brown (4);
Mihaela Ilovan (5);
Kelsey Rubin-Detlev (6)

(1) Newcastle University, UK;
(2) University of Waterloo,
Canada;
(3) Bucknell University, USA;
(4) University of Guelph, Canada;
(5) University of Alberta, Canada;
(6) University of Southern
California, USA

Keywords

TEI XML, Web Editor, Linked Open Data, TEI
Authority Files, Named Entities

Abstract

LEAF-Writer is a popular, free, web-based, semantic editor for Text Encoding Initiative (TEI) and Linked Open Data (LOD) annotation. While LEAF-Writer is available in a standalone version that anyone can use ([LEAF-Writer Commons](#)), it is also a crucial component of the larger [LEAF-VRE](#) environment. LEAF-Writer Commons runs entirely in your browser and files are saved to your own GitHub repositories (or downloaded locally), with no data stored on LEAF servers.

LEAF-Writer provides an easy text-first editing interface that encourages the encoder to focus on adding semantic markup, scholarly notes, and identifying named entities. You can choose to edit in the tags-off view, show tags, or edit the

underlying TEI XML. LEAF-Writer includes schema-constrained context-sensitive tagging and validation using out-of-the-box popular TEI customizations, or use your own custom TEI project schemas (with your own CSS). One of LEAF-Writer's most important features is its built-in support for named entity linking. This enables tagging names of people, places, organisations, or works and associating these (with both TEI markup and LOD) to recognised authorities (such as VIAF, Wikidata, DBpedia, Getty, Geonames, GND and LINCIS). LEAF-Writer can also generate LOD Annotations from already tagged XML references.

Recently, the LEAF team, in collaboration with the [CatCor](#) (Correspondence of Catherine the Great) project has introduced functionality for adding web-accessible project-specific TEI Authority files for particular entity types. (e.g. a TEI 'personography' file containing a `<listPerson>` for person entities, `<listPlace>` for place entities, etc.). This exciting new feature is for many TEI projects that want to do look-ups using their own project-specific '*-ography' authority files.

LEAF-Writer was introduced at TEI2022 and we have run workshops and presented new features to the TEI community in the years since. For this conference, we will concentrate on these 'new territories' of project-specific TEI Authority Files for named entity lookups.

Managing lexical complexity

ODD chaining for Arabic dialect dictionaries

Veronika Engler (1);
Karlheinz Mörrth (1);
Stephan Procházka (2);
Michaela Rausch-Supola (1);
Daniel Schopper (1)

(1) Austrian Academy of
Sciences, Austria;
(2) University of Vienna,
Austria

Keywords

lexicography, corpora, linguistics, data
modelling, ODD chaining

Abstract

The *Vienna Corpus of Arabic Varieties (VICAV)* is a language documentation platform hosting a varied collection of digital language resources. From its beginnings, it has adopted a text-oriented approach that has been grounded in the application of the *Guidelines of the Text Encoding Initiative (TEI)*. The VICAV infrastructure is meant to ensure consistent encoding across projects as well as sustainable creation and publishing workflows. In addition to a bibliography of research literature, typologically similar to what can be found in the *World Atlas of Language Structures* or the *Database of Arabic Dialects*, VICAV offers several types of data: concise descriptions of linguistic varieties, structured lists of linguistic features, sample texts, corpora of unmonitored speech, and dictionaries (Procházka et al. 2015).

The published VICAV dictionaries so far cover five linguistic varieties: Baghdad, Cairo, Damascus, Tunis and Modern Standard Arabic. These dictionaries are all comparatively small, none of them containing more than 8000 entries, and constitute lexical databases with structured lexicographic information (Moerth and Schopper 2021). They are all provided with English translation equivalents (some also include German, French or Spanish translations) and share common encoding conventions.

A recent addition in this series is the *Shawi Dictionary*, which will go online in late 2025. This is the first VICAV dictionary natively encoded in TEI Lex-O, a relatively new initiative within the TEI community, intended as a baseline encoding for lexicographic data (Tasovac et al. 2018) which meanwhile has been adopted by more dictionary projects (Salgado et al. 2019).

Stemming from different projects, all VICAV dictionaries have subtle differences in their encoding requirements. Moreover, there are several other comparable lexicographic resources being developed at the ACDH-CH. To be able to document the differences between those dictionaries and yet keep them compatible, we use the method of ODD chaining, which is the process of deriving ODDs from one another (Pernes et al. 2017). In our case, we proceeded from the *TEI Lex-O ODD* as the primary source and modified it to create the *ACDH-CH generic-dict-schema*, which serves as the baseline ODD for various VICAV and other dictionaries worked on at the ACDH-CH. Since, for example, TEI Lex-O does not make prescribe the macrostructure of a dictionary, it is in this schema that we define that in our dictionaries examples are not embedded within entries but kept in a separate `<div>` in order to allow them to be re-used in different contexts. In a third step, the SHAWI dictionary ODD is derived from the *generic-dict-schema* and provides specifications adapted to the needs of the SHAWI project. Unlike in other dictionaries, the lexical profiles of tribes and their geographic context play an important role. To incorporate this information, we extend *generic-dict-schema* to allow typed `<name>`

elements within `<usg>` – a construct which is not needed in the other dictionaries.

In our paper, we will present the dictionaries involved and our general mechanism next to discussing pros (modularization; specificity) and cons (complexity both in terms of processing overhead and modelling) of having selected this relatively complex route.

References

- Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. WALS Online (v2020.4) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13950591> [Available online at <https://wals.info>]
- Magidow, Alexander. 2015. Database of Arabic Dialects. [Available online at <http://database-of-arabic-dialects.org/>]
- Moerth, K. and D. Schopper (2021) "VICAV 3.0: Zooming in on Lexical Resources", in C. Katsikadelis, M. Sellner and M. Gassner (eds.) *Digital Lexis, and Beyond. Selected Papers from the Workshop „Digital Lexis, and Beyond“ 45th Austrian Linguistics Conference 2019*. Wien: Verlag der ÖAW.
- Pernes, S., L. Romary and K. Warburton (2017) "TBX in ODD: Schema-agnostic specification and documentation for TermBase eXchange", in F. Frontini, L. Grois, Š. Vintar and F. Khan (eds.) *Proceedings of the Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*.
- Procházková, S. and K. Moerth (2015) "The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects", in M. Al-Hamad, R. Ahmed and H. Aloui (eds.) *Lisan Al-Arab: Studies in Contemporary Arabic Dialects, Proceedings of the 10th International Conference of AIDA*, Qatar University 2013. Vienna: LIT Verlag: 209-218.
- Salgado, A., R. Costa, T. Tasovac and A. Simões (2019) "TEI Lex-O In Action: Improving the Encoding of the Dictionary of the

Academia das Ciências de Lisboa", in *Electronic lexicography in the 21st century. Proceedings of the Electronic lexicography in the 21st century (eLex 2019)*: 417-433. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf.

Tasovac, T., L. Romary, P. Banski, J. Bowers, J. de Does, K. Depuydt, T. Erjavec, A. Geyken, A. Herold, V. Hildenbrandt, M. Khemakhem, B. Lehečka, S. Petrović, A. Salgado and A. Witt (2018) *TEI Lex-O: A baseline encoding for lexicographic data*. Version 0.9.3. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILexO/TEILexO.html>. [Accessed 2025-03-27]

TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. [last modified 24.01.2025]. [Available online at <https://www.tei-c.org/P5/>]

Multimodality and Minimal Publishing

TEI, MEI, and more in 19th-Century Music Treatises

Torsten Roeder;
Jana Klinger;
Felicita Stickler;
Corinna Keupp;
Fabian Moss

Julius-Maximilians-Universität
Würzburg, Germany

Keywords

Digital Scholarly Edition, Text and Music,
Minimal Publishing, Multi-Modality

Abstract

Introduction

While presentation methods for the digital edition of either text or music notation can both rely on established tools for TEI or MEI presentation, digital editions that constantly alternate between modalities—e.g., in the context of music theory treatises, textbooks, music journals, composers' correspondence, and editions of sheet music with accompanying peritexts—face significant challenges when combining the related technologies.

This paper builds on a bachelor's thesis dealing with source materials from the project *Digitizing the Dualism Debate* (cf. Moss/Bavaud et al. 2021; Moss/Köster et al. 2021; Moss/Nápoles López et al. 2022). It also connects to the recently reactivated TEI Music SIG, which aims to harmonize and integrate music encoding with the practices of the Text Encoding Initiative

(TEI), focusing particularly on the transitional areas between the two technological stacks. The overarching goal is to offer one or more accessible and sustainable technological solutions for publishing digital editions that include music notation. This approach is intended to support projects without institutional infrastructure for digital editions.

The integration of text and music notation initially appears to be a topographical issue of text presentation. From a text encoding perspective, the problem can be addressed by integrating graphical representations of music notation as digital images. However, from a music encoding perspective, this approach is insufficient, as it prevents the formulation of relationships between text and music notation within the data (cf. Roeder/Moss/Köster 2023). Therefore, the objective is to model and encode both text and music with comparable structural and semantic depth, and to host it within a long-term sustainable environment. While the presentation could integrate scanned or vectorized music notation graphics alongside embedded MEI data, the preferred approach is to generate the presentation directly from the data without intermediate formats (while providing a fallback solution with embedded graphics if necessary).

To illustrate this approach, we propose using two tools closely associated with TEI and MEI: Verovio for rendering music notation and CETEicean, a lightweight tool for TEI processing. Both operate on JavaScript and require only a standard web server, as TEI and MEI are both processed client-wise.

1. Representation: Organizing and Modeling the Data

Due to the different encoding standards, it is necessary to decide how to organize text and music notation data at the file level. To avoid potential namespace conflicts between frameworks, separating them into distinct files has proven effective. This does not necessarily mean generating one TEI file and numerous MEI files; MEI data could also be consolidated into a single file and referenced section by section using IDs.

For example, a treatise with alternating text and musical examples could use a primary TEI file for the text and a single MEI file containing all musical excerpts, each tagged with unique IDs for cross-referencing. This approach reduces complexity while maintaining semantic depth.

2. Presentation: Rendering the Data

Depending on the chosen data organization, the rendering process must be arranged for the browser. With the selected combination of CETELcean and Verovio, HTML and SVG code are generated live using JavaScript. This requires precise timing: a music notation section can only be rendered by Verovio once the corresponding HTML element has been generated by CETELcean. To achieve this synchronization, the frameworks are orchestrated via monitoring. When CETELcean processes an XML element referencing a musical excerpt, it temporarily hands over control to Verovio to render the notation dynamically. This interplay ensures a seamless user experience without relying on pre-generated images or static files.

3. Stabilization: Hosting Environment and Citable Archiving

Dynamic site digital editions are often precarious resources, as they often depend on complex server environments. Even standard solutions (e.g. with TEI Publisher) require regular maintenance, which projects and their institutions are not able to guarantee in long terms. For rendering components such as CETELcean and Verovio, however, simple hosting solutions for static sites like GitHub Pages or any another web server are sufficient (cf. Cayless/Viglianti 2018). GitHub offers advantages such as temporal organization of data and code into release sequences. Additionally, GitHub can be linked to Zenodo, enabling the repository and its individual releases to obtain DOIs and ensuring that each release is archived independently of GitHub on Zenodo. This combination allows projects to achieve both low-cost deployment and academic curation, compliant with the

endings principles (cf. The Endings Project 2023). For instance, an edition hosted on GitHub Pages could be continuously updated, while its key versions remain permanently accessible via Zenodo.

Outlook

The Würzburg project *DigiMusTh: Aufbau einer offenen digitalen Sammlung historischer musiktheoretischer Texte aus dem deutschsprachigen Raum anhand von Beispielen aus dem 19. Jahrhundert* ("Development of an open digital collection of historical music theory texts from the German-speaking world based on examples from the 19th century") will adopt this new presentation model in 2025 using the described approach. Simultaneously, this approach will serve as an example for the TEI Music SIG, showcasing its potential for other interdisciplinary projects. The next step involves modeling the dense relationships between text and music notation and integrating these into the presentation. It will also be discussed and presented how to include further rendition methods for other modalities, such as mathematical formulas in MathML or diagrams in SVG.

References

- Cayless, H., Viglianti, R. (2018). CETELcean: TEI in the Browser. In Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21. <https://doi.org/10.4242/BalisageVol21.Cayless01>.
- CETELcean (2024). Release 1.9.0 (Jan 10), <https://teic.github.io/CETELcean>.
- The Endings Project (2023). Endings Principles for Digital Longevity, Version 2.2.1, 2023-03-03. <https://endings.uvic.ca/principles.html>.
- Moss, F. C., Bavaud, F., Métrailler, C., Femminis, M., Köster, M. (2021). Digitizing the Dualism Debate. <https://dcmlab.github.io/ddd/>.

- Moss, F. C., Köster, M., Femminis, M., Métrailler, C., & Bavaud, F. (2021). Digitizing a 19th-Century Music Theory Debate for Computational Analysis. In M. Ehrmann, F. Karsdorp, M. Wevers, T. L. Andrews, M. Burghardt, M. Kestemont, E. Manjavacas, M. Piotrowski, & J. van Zundert (Eds.), *CHR 2021: Computational Humanities Research Conference*, November 17-19, 2021, Amsterdam, The Netherlands (pp. 159-170). CEUR. http://ceur-ws.org/Vol-2989/short_paper31.pdf.
- Moss, F. C., Nápoles López, N., Köster, M., & Rizo, D. (2022). Challenging sources: A new dataset for OMR of diverse 19th-century music theory examples. In J. Calvo-Zaragoza, A. Pacha, & E. Shatri (Eds.), *Proceedings of the 4th International Workshop on Reading Music Systems (WoRMS 2022)* (pp. 4-8). <https://sites.google.com/view/worms2022/proceedings>.
- Roeder, T., Moss, F. C., Köster, M. (2023). Musio-Text Interlinking as a Challenge for Digital Encodings of Music-Theoretical Writings, TEI/MEC 2023. <https://teimec2023.uni-paderborn.de/contributions/124.html>.
- Verovio (2024). Version 4.4 (Dec 17). <https://www.verovio.org>.

Navigating and Processing Data from the TEI with XPath and XSLT

Elisa Beshero-Bondar (1);
Martina Scholger (2);
Patricia O'Connor (3)

(1) Penn State Erie, USA;
(2) University of Graz, Austria;
(3) Maynooth University, Ireland

Keywords

XPath, XSLT, pull processing, data visualization, internationalization

Abstract

TEI markup provides structures that are particularly useful for processing data beyond what we can do with so-called “plain text”.

Our workshop teaches the pull-processing of data from XML/TEI with simple, reusable XSLT templates to represent in simple TSV/CSV, HTML tables/charts, and (if time!) simple SVG graphics.

Knowing how to locate and explore data in your encoding can help to learn how to work with TEI and XML generally. This half-day workshop is designed for people who have some experience with TEI and seek to learn how to work with XML markup for analysis and research. Participants will gain a working, practical knowledge of the query language XPath and the transformation language XSLT, and learn how these can help to reduce reliance on software, packages and plugins that may become obsolete without warning. Further, XSLT’s functional programming can serve as a way of articulating research questions around a document data model expressed in XML.

The emphasis of our workshop is “pull-processing”, that is, extracting data and metadata from markup documents for analysis, as opposed to providing the reading view of a digital scholarly edition. Markup in documents supplies structures and contexts that are especially useful for processing data, beyond what we can do with so-called “plain text”. We will demonstrate some basic XPath navigation and calculation functions, and then show how XPath is applied in XSLT templates to address specific nodes that hold data of interest for visualization.

We will process TEI documents composed in various languages represented by our workshop members’ projects, to show that the code we write is transferable to multiple projects across language and cultural borders.

Participants will learn how to “pull” data from TEI and output text formats required for simple online tools, where the structure of the output data is transferable to many different online calculation programs and amenable to statistical processing. During the workshop we will produce some simple structured documents for storing, sharing, and visualizing data: HTML lists and tables as well as plain text tabulated data (CSV or TSV files), and (if we have time) simple SVG bar or line graphs.

We hope to process some participant-supplied XML before, during, and after the workshop. We will carefully document the XSLT that we supply during the workshop to assist participants with revising and adapting the code to their own projects.

The workshop material - including documentation, exercise examples and solutions - will be made available in a publicly and permanently accessible GitHub repository.

New features of Scholarly XML, an Open Source Visual Studio Code Extension for TEI encoding

Raffaele Viglianti

University of Maryland, USA

Keywords

tools, minimal, encoding support, schematron

Abstract

First released in 2020, Scholarly XML is an open source extension for Visual Studio Code (VS Code) <https://marketplace.visualstudio.com/items?itemName=raffazizzi.sxml>. It has provided a lightweight and easy to use solution for simple TEI encoding work, such as RELAX NG validation and schema-aware suggestions. The extension has nearly 15,000 installs and has been awarded the 2024 TEI Rahtz Prize for Ingenuity, which has prompted a new phase of development. This poster will re-introduce this tool to the TEI community, showcase its latest features, and provide an opportunity to shape its future development.

Scholarly XML was originally developed for teaching purposes, at a time when there was no clear free and easy to set up choice for introducing students with minimal technical skills to TEI encoding. Since then, the extension has been successfully used for encoding work in research projects, testing its

applicability outside of a pedagogical context. While the landscape has changed little, the very popular Red Hat's "XML" extension for VS Code (<https://marketplace.visualstudio.com/items?itemName=redhat.vscode-xml>) has met many more of the requirements needed for introductory TEI work. In particular, since v0.15.0 released in March 2021, it no longer requires Java installed on users' systems, which poses a major limitation for adoption in the classroom and by less technically inclined encoders. RELAX NG support was moreover introduced in Red Hat XML v0.22.0 released in November 2022. Scholarly XML, however, aims to cover requirements of the TEI community that are not fully met by a more generic extension.

The latest phase of development, with a new release (v0.4.0) planned before the conference in Kraków, includes support for schematron validation (with contributions by Joey Takeda, University of British Columbia, and Maryland undergraduate students Evan Henkle and Eshan Singh) and XInclude support for full document inclusion (with contributions from Eshan Singh).

	Red Hat XML (v0.28.0)	Scholarly XML (v0.4.0rc)
RELAX NG validation	Yes	Yes
XSD validation	Yes	No
Independent from additional software	Java required to run extensions to the base XML features	Yes
Schema-aware suggestions	Yes	Yes
Support for RELAX NG Annotations	Yes	Yes
Schematron	No (there is a third-party extension that requires Java)	Yes
XInclude	Partial, no errors reported on parent XML; partial xpointer support.	Full document inclusions (attributes href and parse="xml"); errors reported on parent document
XML Catalogs	Yes	No

New Territories for (Very) Old Language: DiaCorPolL

Automated Compilation of the Diachronic Corpus and Dictionary of Latin in Poland

Krzysztof Nowak;
Iwona Krawczyk;
Jagoda Marszałek

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

Keywords

corpus linguistics, HTR technologies, Latin,
diachronic research

Abstract

The history of Latin in Polish lands extends from the threshold of the second millennium to at least the end of the 18th century (Axer 2004). Despite significant scholarship on individual Latin authors and epochs, a synthetic approach to the historically variable functions and forms of Latin in Polish territories remains absent, largely due to the lack of appropriate research tools.

The project *DiaCorPolL: Diachronic Corpus of Latin in Poland* aims to create an electronic text database enabling in-depth empirical research on Latin language from the Middle Ages through the Renaissance and Baroque periods until 1800. In this paper, we discuss (1) the project's assumptions, (2) methods of data acquisition, and (3) the development of an LLM-based system for automatic drafting of dictionary entries.

The diachronic corpus aims to encompass all Latin texts created in Polish territories from approximately 1000 to 1800, as

well as Latin translations and works by foreign authors published in Poland, and Latin works by Polish authors published abroad. The selection reflects the diversity of Latin functions, encompassing literary, scientific, and pragmatic texts. Text selection is based on syntheses of Polish literature, bibliographies, and library databases and catalogs (see Figure 1).

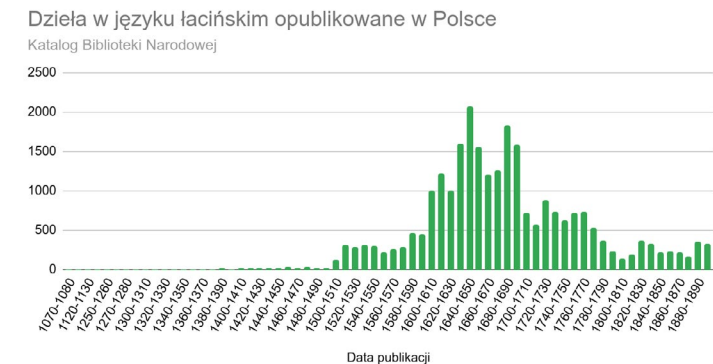


Fig. 1: Number of Latin works published in Poland from the Middle Ages to 1900 (Source: <https://data.bn.org.pl>)

In acquiring linguistic material, we utilize OCR and HTR tools to process contemporary editions, manuscripts, and old prints from digital libraries (see Figure 2). Existing segmentation and recognition models are adapted specifically for processing Polish resources. TEI Publisher is used to present both the digital facsimile and the OCR and HTR output.

The project aims to enable not only comprehensive diachronic research on Latin language in Poland, including its complex relationships with Polish (Bronikowska and Kryńska 2020), but also to create lexical resources, including dictionaries. However, funding limitations and the sheer volume of texts necessitate automatic treatment of corpus data and extensive use of LLMs, potentially transforming how we document and analyze

semantic evolution in historical languages. We present the results of initial experiments aimed at automatic dictionary entry drafting using a streamlined LLM-based system.

Finally, we demonstrate how automatic methods help in leveraging existing TEI XML and Semantic Web resources to enrich our knowledge of Latin use across centuries and locations, primarily through the electronic corpus (Nowak 2024) and dictionary of Polish Medieval Latin (Nowak 2014).

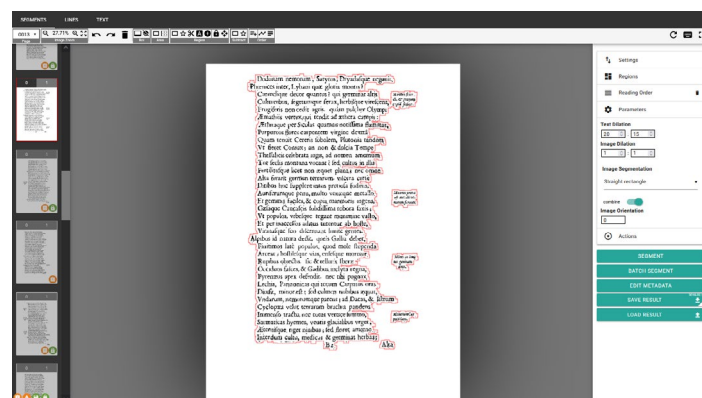


Fig. 2: Region recognition

References

- Axer, Jerzy, red. 2004. *Łacina jako język elit*. Warszawa: OBTA: DiG.
- Bronikowska, Renata, and Katarzyna Kryńska. 2020. „Łacina w KorBie. Użyteczność Elektronicznego Korpusu Tekstów Polskich XVII i XVIII wieku dla filologa neolatynisty”, *Polonica* 40. <https://doi.org/10.17651/POLON.40.8>.
- Nowak, Krzysztof. 2014. 'The eLexicon Mediae et Infimae Latinitatis Polonorum. The Electronic Dictionary of Polish

Medieval Latin'. In *The User in Focus. Proceedings of the XVI EURALEX International Congress: 15 - 19 July 2014, Bolzano/ Bozen*, edited by Andrea Abel, Chiara Vettori, and Nataschia Ralli, 793-806. Bolzano: EURAC Research. http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

Nowak, Krzysztof. 2024. „Korpusy językowe. Budowa, wykorzystanie, potrzeby”. W *Człowiek twórcą historii*, zredagowane przez Cezary Kukło i Wojciech Walczak, 5;1: Warsztat nowoczesnego humanisty historyka na progu XXI w.: 51-72. Białystok: Uniwersytet w Białymstoku.

New Territories for Digital Lexicography

Dictionary of Polish Dialects meets TEI

Dorota Mika;
Krzysztof Nowak

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

Keywords

electronic lexicography, linguistics,
dialectology

Abstract

Regional varieties of language constitute vital cultural heritage, with their documentation and preservation being increasingly urgent tasks. This paper introduces the digital edition of the Dictionary of Polish Dialects, presenting our workflow from print conversion to web publication.

The Dictionary of Polish Dialects (Karaś et al. 1979) is a monumental multi-volume work compiled by generations of scholars from the Institute of Polish Language of the Polish Academy of Sciences. It contains unique 19th and 20th-century materials from Poland and neighboring regions including Lithuania, Belarus, Ukraine, the Czech Republic, Slovakia, Romania, and Hungary. Despite ongoing development, its paper-only format has restricted access and data exploration possibilities. Digitization has become essential both for preserving linguistic

heritage and enabling new research into the language, history, and culture of Central and Eastern Europe.

Our process began with OCR preprocessing and conversion to temporary XML files, followed by implementation of a complex XProc pipeline with multiple XSL stylesheets to maximize lexicographic data encoding. We then developed TEI-conformant encoding (The TEI Consortium 2025) that balanced editorial idiosyncrasies with requirements for harmonizing linguistic descriptions across resources (Chiarcos et al. 2013). Particularly challenging aspects included bibliographic references, non-standard orthography, phonetic transcription characters, and extensive geographic information. The resulting TEI XML is converted via XProc to tabular format and exported to RDF through custom mapping.

The project ensures interoperability through TEI encoding standards and extensive use of existing ontologies and shared vocabularies. We offer dual access points: a TEI Publisher instance for text-centered exploration (see Table 1) and a semantic wiki (Bon and Nowak 2013) enabling complex spatial queries that leverage the dictionary's rich geographic data.

ALKIERZYK

Formy: Alkierz *Huszcz* biał-podl ; - aukl' iżyk || alkiżyk ||- Domaniewek łącz *PJPAN* XXXIII 17 ; „uklnażyk || alkiżyk *Kramsk* koniń ; alkiży *Domaniewek łącz* *PJPAN* XXXIII 17 ; - jaŋkl' iżyk || alkiżyk ||- jw .

Znaczenie:

'mały alkierz, izdebka o różnym przeznaczeniu':

→ Pokoik jadalny, dość często nazywany „alkierzykiem” (jeżeli jest jeszcze więcej mieszkańia) *Pabianice Wzrost* IV 827 ;

→ ćce do alkiżyka, co fluńi stojce! *Huszcz* biał-podl ;

→ bapka miśkaya fitym alkiżyku; zaprosiły nożyobuz do aukl' iżyka i ugościły *Kramsk* koniń ;

- 'sypialnia' : *Domaniewek łącz* *PJPAN* XXXIII 17 ; *Wierchjedlina* sokół .

WM

Fig. 1: TEI-Publisher-based Web Edition

References

- Bon, Bruno, and Krzysztof Nowak. 2013. 'Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki'. In *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, edited by Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets, and Maria Tuulik, 407-20. Tallinn - Ljubljana: Trojina, Institute for Applied Slovene Studies; Eesti Keele Instituut. https://ellex.link/ellex2013/wp-content/uploads/eLex2013_28_BonNowak.pdf.
- Chiarcos, Christian, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. 'Towards Open Data for Linguistics: Linguistic Linked Data'. In *New Trends of Research in Ontologies and Lexical Resources*, edited by Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy, 7-25. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-31782-8_2.
- Karaś, Mieczysław, Jerzy Reichen, Joanna Okoniowa, and Renata Kucharzyk, eds. 1979. *Słownik Gwar Polskich*. Kraków: Instytut Języka Polskiego PAN.
- The TEI Consortium. 2025. 'The TEI Guidelines. P5 Version 4.9'. Text Encoding Initiative Consortium.

ODD-API

A REST Interface for Programmatic Use of ODD Schema Definitions

Peter Stadler (1);
Anne Ferger (1);
Johannes Kepper (1);
Raffaele Viglianti (2)

{1} Paderborn University, Germany;
{2} University of Maryland, USA

Keywords

API, ODD, REST, eXist, XQuery, JSON:API

Abstract

The ODD-API provides a REST interface that enables programmatic access to ODD schema definitions [Viglianti 2019]. In contrast to traditional approaches, where ODD documents must be manually processed with XML tools, this API offers a standardized method to query information about elements, attributes, classes, and modules of a schema via HTTP.

By using the JSON:API format [Yehuda et al. 2022], the API is particularly user-friendly and compatible with any programming language that supports HTTP requests. The API is documented according to the OpenAPI specification [OpenAPI Initiative 2024], allowing applications built on the API to be developed simply and efficiently without having to understand the complex ODD structure nor XML.

Originally, the API (in version 1) was developed as a backend for the [MEI Profile Drafter](#) (a lightweight ODD editor for MEI customizations), but it had a generic approach from the beginning and also allowed the use of TEI ODD schemas. This generic approach

will be further expanded with version 2 of the API, making the ODD-API usable as a central hub for ODD schema definitions in general, creating various new application possibilities. The API can thus serve as a foundation for developing tools to compare ODD schemas (or different versions of a schema), to edit ODD documents, to visualize dependencies between ODD schemas, for learning platforms to teach ODD knowledge, or dynamic documentation systems to complement existing static websites.

Implementation

As an initial implementation for the ODD-API scheme, we relied on the eXist database and XQuery. Various ODD schema definitions (e.g., `teip5subset.xml`) are stored in defined database collections by type and version (e.g., `/db/apps/odd-api/data/tei/v4.9.0`); additional ODD schema definitions may be loaded by dedicated endpoints. Currently, the following endpoints for GET requests are implemented:

- `/v2/schemas`: returns a list of all available schemas in JSON:API format
- `/v2/{schema}`: returns a list of all versions of a schema in JSON:API format
- `/v2/{schema}/{version}`: returns the version of a schema as "Compiled TEI ODD" in XML format
- `/v2/{schema}/{version}/elements`: returns a list of all elements in JSON:API format
- `/v2/{schema}/{version}/attributes`: returns a list of all attributes in JSON:API format
- `/v2/{schema}/{version}/classes`: returns a list of all classes in JSON:API format
- `/v2/{schema}/{version}/modules`: returns a list of all modules in JSON:API format

The response for the endpoint `/v2/tei` (on a local test instance at `http://localhost:8080`) in JSON:API format looks like this:

```
{
  "data": [
    {
      "type": "versions",
      "id": "version_tei_4.6.0",
      "attributes": {
        "version": "4.6.0"
      },
      "links": {
        "self": "http://localhost:8080/v2/tei/4.6.0"
      }
    },
    {
      "type": "versions",
      "id": "version_tei_4.8.1",
      "attributes": {
        "version": "4.8.1"
      },
      "links": {
        "self": "http://localhost:8080/v2/tei/4.8.1"
      }
    }
  ],
  "links": {
    "self": "http://localhost:8080/v2/tei"
  }
}
```

According to the JSON:API specification, each object must contain a `type` attribute and an `id` attribute. The latter must be unique within the API, so in this example, the version number is prefixed with the type and the schema name. The pure version number is returned in the `attributes` object.

Application Scenarios and Outlook

A concrete current application scenario for the ODD-API is its use as an information resource for the ODD editor [Roma](#). Roma is to be further developed to allow editing of non-TEI schemas and requires canonical addresses of ODD resources as starting points for customizations. Here, the ODD-API can serve as a central point to provide this information. To enable this in other contexts as well, only the URL of the ODD-API (analogous

to the URL of the TEIGarage) would need to be specified as a parameter in the Roma code. Private Roma instances could then address any other ODD-APIs with any ODD schema definitions without having to adapt the source code of Roma or the ODD-API.

Beyond that, it would also be possible to open the ODD-API for PUT requests (and potentially also POST and PATCH) and thus also allow adding and editing ODD schemas. This would enable even stronger dynamization of the ODD-API and could be used, for example, in automated pipelines. However, this would then require appropriate authentication and authorization as well as guidelines for handling additions by third parties.

Finally, it would also be possible to develop an ODD package manager, i.e., an “npm for ODD schemas.” This package manager could have an integrating effect for the ODD community by highlighting the diversity and possibilities of ODD schemas on the one hand, while promoting standardization on the other. The API defined here would provide the underlying engine for such a hypothetical package manager.

References

- Katz, Yehuda, Dan Gebhardt, Gabe Sullice, and Jeldrik Hanschke. 2022. “JSON:API Specification.” <https://jsonapi.org/>
- OpenAPI Initiative. 2024. “OpenAPI Specification v3.1.1.” <https://spec.openapis.org/oas/v3.1.1>
- Viglianti, Raffaele. 2019. “One Document Does-it-all (ODD): a language for documentation, schema generation, and customization from the Text Encoding Initiative.” Presented at Symposium on Markup Vocabulary Customization, Washington, DC, July 29, 2019. In Proceedings of the Symposium on Markup Vocabulary Customization. Balisage Series on Markup Technologies, vol. 24 (2019). <https://doi.org/10.4242/BalisageVol24.Viglianti01>.

Parallel Editing in TEI

The Case of *Regestra 1561*

Łukasz Poznański;
Bogumił Szady

Humanities Research Data
Lab, The John Paul II Catholic
University of Lublin (KUL), Poland

Keywords

TEI, parallel edition, historical documents,
linked data, TEI Publisher

Abstract

This paper presents *Regestra 1561*, a parallel digital edition of two manuscript copies of a clerical tax register from the diocese of Poznań. Both documents, preserved in the Central Archives of Historical Records in Warsaw, were compiled following the 1561 provincial synod. They record contributions from local clergy, including names, ecclesiastical benefices, places, and amounts of taxes owed to the Crown.

As part of the *Regestra 1561* project, both a traditional full-text print edition and a digital edition were prepared. The digital edition was designed to enable a side-by-side comparison of both witnesses. Encoded in TEI-XML, it captures the full text of each manuscript and aligns corresponding sections, allowing users to explore both similarities and differences in content and structure.

Creating such an edition posed challenges: the manuscripts diverge in layout and pagination, requiring a parallel markup

system that preserves each witness's structure while linking related segments. The edition also includes a TEI-encoded critical apparatus and detailed references to people and places.

The edition is presented in TEI Publisher with customized views supporting synchronized navigation of both manuscripts. Thanks to IIIF integration, the texts appear alongside digitized folios. Differences between the versions are automatically highlighted.

The edition also links to a structured database hosted on wiki.kul.pl, which offers semantic context for named entities. To enable searching and filtering via TEI Publisher, the contents of this external database were exported and integrated into the TEI source files. This allows users to query and explore the edition using enriched metadata based on normalized person and place records.

The project illustrates how TEI-based parallel editing, combined with linked data and visual tools, can offer an interactive presentation of early modern administrative documents.

Processable and Computable Media in TEI

Usecases and Strategies

Torsten Roeder;
Tomasz Shtohryn;
Yannik W. Herbst

University of Würzburg, Germany

Keywords

paper forms, e-literature, digital
correspondence, program code,
punchcards

Abstract

Beyond Readability?

In everyday usage, "text" typically refers to physical or electronic inscriptions – material or digital – that are intended for human reading. Our notion of textual heritage predominantly encompasses classical forms such as literature, correspondence, journals, and newspapers, all seemingly bound to natural language. The TEI Guidelines offer extensive support for encoding such textual complexity, accommodating a wide range of human-created text types. Although such texts are, in principle, amenable to processing by computational rulesets or human interpretation, this kind of processing is not usually the primary focus. This prompts a set of critical questions: Is "text" – and by extension, the TEI – necessarily bound to human natural

language? Conversely, are there classes of textual heritage that require mechanical or rule-based processing either before, during, or after human or machine reading? If so, how can these be encoded within the existing TEI framework, and what expansions might be necessary? Furthermore, how can the rulesets for such processing themselves become integral to the preservation of cultural heritage?

Examples

The questions raised may appear broad, if not esoteric, thus it is appropriate to illustrate them with concrete examples. Consider a multiple-choice examination sheet completed by a student and graded by a teacher. The form itself encodes an implicit ruleset that governs how the examination is to be completed and assessed – thus undergoing multiple stages of processing to produce a graded result.

Another example stems from early digital literature: HyperCard-based works require specific computational environments (e.g., pre-OS X Macintosh systems) for proper functionality. The user's interaction with these works is defined by both the hardware and software, raising the question of whether essential aspects of historical digital interfaces can be preserved through TEI encoding (cf. Ensslin 2007 for a more general discussion on early hypertext fiction).

Similarly, early digital periodicals such as "diskmags" – early software magazines distributed on floppy disk – often employed custom character sets and graphical conventions that defy standard extraction and representation methods. Here, knowledge of the original processing mechanisms is essential for reconstructing the intended presentation of text, images, and audio content (cf. Shtohryn 2025).

Emails, a contemporary form of correspondence, further complicate notions of text (cf. Beshero-Bondar and Bauman 2024). Beyond metadata such as sender and recipient information,

the software environment critically influences the creation, display, and transmission of emails. Variations between client applications and server-side modifications introduce layers of textual genetics and performance dynamics that are relevant for scholarly documentation.

Finally, historical program code published in print media ("paperware") exemplifies the hybrid nature of text and process (cf. Höltgen 2014). Such code can be "read" and executed by both humans and machines. Punchcards, featuring both computer-readable perforations and human-readable printed text, reinforce the idea that digitality often straddles material and electronic forms (cf. Feichtinger 2023).

Perspectives

These examples collectively demonstrate that distinctions between "digital" and "material" documents are secondary when considering a document's intentional processability. While this poster centers primarily on born-digital heritage, it recognizes the blurred boundaries between digital and material media. Recent TEI developments – such as the introduction of the `<post>` element by the special interest group for "Computer-Mediated Communication" in 2025 (cf. TEI 2025) – illustrate growing recognition of these evolving forms. Given the increasing volume of born-digital cultural heritage and its vulnerability to hardware decay, software obsolescence, corporate restrictions, and political censorship, it is urgent to develop strategies for preserving processable and computable forms of text.

This poster will present the aforementioned examples to initiate a discussion on extending TEI's capabilities in this area. We aim to establish a new TEI Special Interest Group dedicated to processable and computable media – recognizing that the traditional notion of "text" may not be sufficiently inclusive in this context. Our initial focus will be on utilizing existing guidelines, proposing semantic extensions where necessary, and, where appropriate, developing new attributes or elements.

References

- Beshero-Bondar, Elisa; Bauman, Syd (2024): Can we apply the new CMC chapter to the TEI Listserv Archives? An experiment with TEI for Correspondence and Computer-Mediated Communication, TEI Conference 2024. <https://zenodo.org/records/13957666>
- Ensslin, Astrid (2007): Canonizing Hypertext. Explorations and Constructions.
- Feichtinger, Moritz (2023): Annotation, Simulation und Analyse eines historischen Datenbanksystems. In: Burghardt, M. & Weiß, C. (eds.): Lecture Notes in Informatics (LNI), Gesellschaft für Informatik.
- Höltgen, Stefan (2014): Humanities of the Digital. Philologische Perspektiven auf Source Codes als Beitrag einer computerarchäologischen Knowledge Preservation. In: Bartelmus, M./Nebrig, A. (eds.): Digitale Schriftlichkeit. Programmieren, Prozessieren und Codieren von Schrift, p. 207-229.
- Shtohryn, Tomash (2025): Semi-automatisierte Erzeugung eines Textkorpus von deutschsprachigen Diskettenmagazinen für das Heimcomputersystem Commodore 64 im TEI-Format, Universität Würzburg. <https://doi.org/10.25972/OPUS-40282>
- TEI (2025): Computer-mediated Communication. In: TEI: Guidelines for Electronic Text Encoding and Interchange, P5 Version 4.9.0, 24th January. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CMC.html>

Reconstructing the Experience of a Historian from the Meiji (明治) Period

Attempts to Encode Kengo Murakawa (村川堅固)'s Diary in TEI

Jun Ogawa (1);
Yuko Fukuyama (2);
Shozo Matsutani (3)

(1) The University of Tokyo, Japan;
(2) Waseda University, Japan;
(3) Japan Society for the
Promotion of Science, Japan

Keywords

East Asian materials, diary, writing mode, historiography

Abstract

In this paper, we discuss the Murakawa Collection, a compilation of documents related to Kengo Murakawa (1875-1946), a significant figure in Japanese academia, particularly in the field of Western ancient history [1]. It encompasses various documents such as postcards, lecture notes, and household accounts. Here, we specifically focus on his diary as a case study for TEI encoding because it enables us to gain insights into his live experiences and thoughts.

As TEI encoding methods for diary documents have already been proposed in previous studies like The Harry Watkins Diary [2] and Shibusawa Eiichi Diary [3], we decide to adopt them to encode the basic structure of our texts. Each entry, typically corresponding to a single date (see Fig. 1), is structured using `<div>` with `@type='entry'`, and physical page breaks are marked with `<pb/>`. Header information such as date, location, and weather, often found at the start of each entry, is encoded with `<dateline>`, while entities like personal names and places are also tagged with `<persName>`, `<placeName>`, etc. as shown in Fig. 2.

In addition to structure and entities, we link text to corresponding images at the page level by adding an `@facs` value to each `<pb/>`, referencing the `<graphic>` element in the `<facsimile>` block. Our encoding also addresses text orientation, a unique feature of our diary which includes both vertical and horizontal text, often requiring the book to be rotated for proper reading, preserved through `@style` for each `<pb/>` [4].

Our project tackles significant issues in historical Japanese handwritten sources, such as text orientation and the integration of Japanese and Western languages. We apply established diary encoding methods to our specific sources and are developing a diary viewer (Fig. 3) using structured TEI texts, advancing digital preservation and analysis of our materials.

Acknowledgement

We are profoundly grateful to Natsuko Murakawa for kindly granting permission to view and digitally preserve the Murakawa family materials discussed in this paper. We also deeply appreciate the detailed insights she provided regarding the contents of these materials.

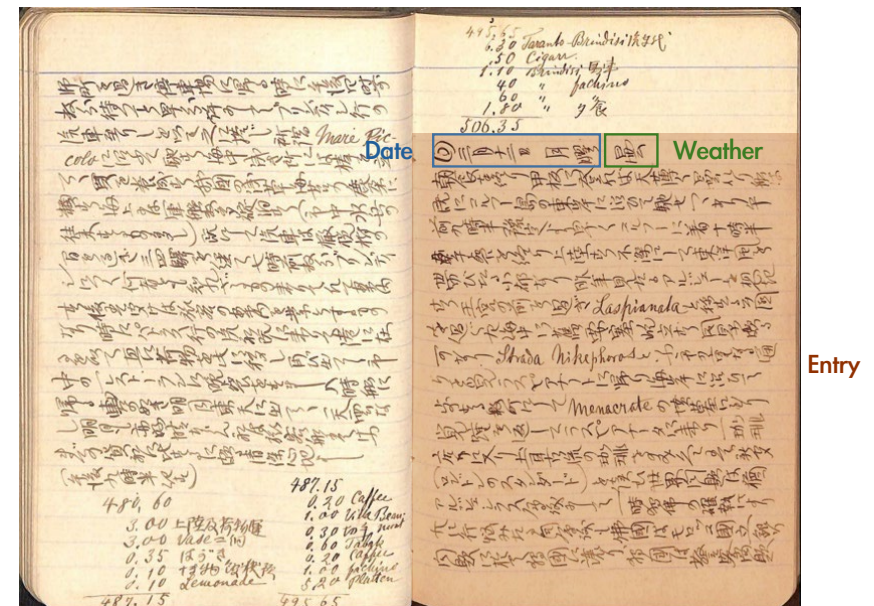


Fig. 1: Basic structure of a single diary entry

```
<div type="entry">
  <pb type="recto" n="57" facs="手帳1_00057" style="writing-mode: vertical-rl; text-orientation: mixed"/>
  <dateline><date when="1884-03-12">三月十二日</date> 月曜 <seg type="weather" and="#cloudy">
    <seg/></dateline>
  <p>
    <lb/>朝食を終り、甲板に登れば、天薄く曇れり。船は
    <lb/>既に<placeName ref="#CORFU">コルファ</placeName>の東岸に沿ひて駛せつあり。<time when="09:00:00">午
    <lb/>前九時半</time>、予定より早く<placeName ref="#CORFU">コルファ</placeName>に着。<time>十時半</time>、
    <lb/>午餐を終り、上陸す。不潔にして東洋風を
    <lb/>帯びたる小都なり。昨年見たる<placeName ref="#ALGI">アルジェ</placeName>を想起
    <lb/>す。王宮の前を通ぎ、<placeName and="#CORFU">スピアナーダ [Laspianata]</placeName>と稱する公園<placeName>
    <lb/>を過ぐ。左海中に旧要塞屹立す。風景頗る
    <lb/>可なり。<placeName and="#CORFU">ニケフォロス通り [Strada Nikephoros]</placeName>といふ。市の重なる通
    <lb/>りを覽、<placeName and="#CORFU">スピアナーダ</placeName>に帰る、海岸に沿ひて
    <lb/>歩む。数町して、<persName and="#CORFU">メネクラテス [Menecrate]</persName>の墳墓に至り、
    <lb/>覽る。踵を返して、<placeName and="#CORFU">スピアナーダ</placeName>に來たり。一咖啡店
    <lb/>に入り、トルコ流の咖啡をのみ、久々に新聞
    <lb/> (ロンドンのスタンダード) を読む。世界問題は猶
    <lb/>アルヘシラス會議にして、一時独仏の確執により
    <lb/>共に行動したる同會議も、仏國は<orgName>モロッコ國立銀行</orgName>
    <lb/>問題に於て獨國に譲り、獨國は警察問題
  </p>
  <pb
    type="verso" n="58" facs="手帳1_00058"
  />
  <lb/>に於て獨國に譲り、十日の會議にて無事
  <lb/>落着すべしとあり。英國の國勢調査の結果
  <lb/>を覽るに大英帝國總人口は四億、母國は四千
```

Fig. 2: Markup example of a selected entry from the diary

◎三月十二日 月曜 曇

朝食を終り、甲板に登れば、天高く曇れり。船は既にコルフー島の東岸に沿ひて駛せつつあり。午後一時、予定より早くコルフーに着。午後二時、午餐を終わり、上陸す。不潔にして東洋風を帯びたる小都なり。昨年見たるアルジェを想起す。王宮の前を通ぎ、スピアナダ〔Laspianata〕と称する公園を過ぐ。左海中に旧要塞屹立す。風景頗る可なり。ニケフォロス通り〔Strada Nikephoros〕といふ。市の重なる通りを覽、スピアナダに降り、海岸に沿ひて歩む。数町にして、メネクラテス〔Menecrate〕の墳墓に至り、覽る。踵を返して、スピアナダに来たり。一咖啡店に入り、トルコ流の咖啡をのみ、久々に新聞（ロンドンのスタンダード）を読む。世界問題は猶アルヘシラス会議にして、一時独仏の確執により共に行動みたる同會議も、仏國はモロッコ國立銀行問題に於て獨國に譲り、獨國は警察問題

58 - verso

に関して獨國に譲り、十日の會議にて無事落着すべしとあり。英國の國勢調査の結果を覽るに大英帝國總人口は四億、母國は四千万余。千八百六十一年以後の四十年間に英領地の面積は四割を増したりとあり。英國未だ老いずと謂ふべし。午後一時、海岸に降り、前に

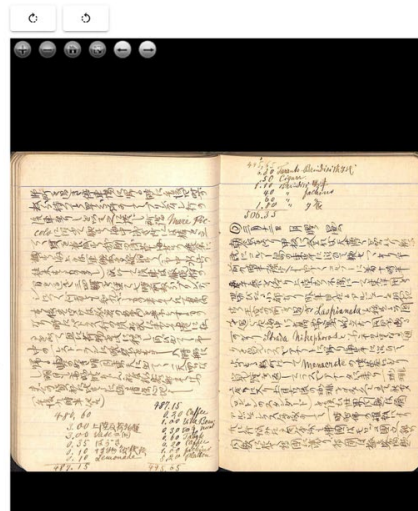


Fig. 3: TEI-based viewer for Murakawa's diary

References

- [1] www.l.u-tokyo.ac.jp. (n.d.). *The History of the Faculty of Letters - Graduate School of Humanities and Sociology / Faculty of Letters, University of Tokyo*. [online] Available at: <https://www.l.u-tokyo.ac.jp/eng/history/history.html> [Accessed 11 Jul. 2025].
- [2] [Umich.edu](http://umich.edu). (2018). Index | The Harry Watkins Diary: Digital Edition | University of Michigan Library Digital Collections. [online] Available at: <https://quod.lib.umich.edu/h/hwatkins/> [Accessed 11 Jul. 2025].
- [3] 渋沢栄一ダイアリー. (2025). 渋沢栄一ダイアリー. [online] Available at: <https://shibusawa-dlab.github.io/app1/> [Accessed 11 Jul. 2025].
- [4] Tei-c.org. (2025). *5 Characters, Glyphs, and Writing Modes - The TEI Guidelines*. [online] Available at: <https://www.tei-c.org/release/doc/tei-p5-doc/ja/html/WD.html#WDWMEG> [Accessed 11 Jul. 2025].

Save it on a Shoe String

Experiences with Project Endings

Constance Crompton

University of Ottawa, Canada

Keywords

sustainable web development,
metadata, search

Abstract

The Humanities offer superpowers to the public: the ability to understand the past, avoid its mistakes, and make evidence-backed decisions about the future (Nussbaum). When used in online knowledge creation and dissemination, Humanities superpowers have the might to stave off an intellectual dark age, provided that we prevent a digital dark age characterized by the disappearance of Humanities data from the Web ecosystems (Barats; Carlin et al 2023a; Crompton 2023; Maron et al.). We know the Web ecosystem is fragile as we have experienced this die-off on the Web before: from its earliest days, the Web was a way to amplify marginalized voices from the past. Indeed, throughout the 1990s, many “scholars invested in early work on race (and class and gender in) the digital humanities insisted on building editions and digital texts as an activist intervention in the closed canon” (Earhart, 317). Many of the earliest digital project content has disappeared from the Web when it in formats that could not be sustained by home institutions (Goddard and Seeman; Jenstad, Holmes

and Huoulak]. This has resulted in the loss in particular of on-line Humanities scholarship, or in partial snapshots of projects captured by third-party services, like the Internet Archive. There is much room to better serve and preserve Humanities projects and to continue to share their vital knowledge in the way we do Humanities monographs without increasing the cost and technological burden to academic libraries and other stakeholders.

This short paper introduces the Lesbian and Gay Liberation in Canada project's experience with the principles and code based developed by the Endings Project at the Humanities Computing Media Centre at the University of Victoria ("Endings"). Lead by the tech lead who first used Endings beyond the original Endings team, we outline the benefit of switching to a static-only version of the site, and the benefits of this approach in the current North American financial and political context.

Standardisation and innovation

The role of TEI in improving innovative potential in digital scholarly editions of correspondences

Agnieszka Szulińska

Polish Academy of Sciences,
Poland

Keywords

digital scholarly editions, scholarly communication, correspondence, letter

Abstract

In this talk, I want to present the first phase of a broader project exploring digital scholarly editions as innovations in scholarly communication and, more specifically, the role of the TEI standard in enhancing their innovation potential.

Innovation, understood as "novel approaches, tools, or workflows that enhance the creation, dissemination, and evaluation of scholarship" (Maryl, Wnuk, Gouzi and Umerle 2024), could become an important point influencing the perception of scholarly digital editing in Poland. Digital editions have been created in more Polish scholarly institutions relatively recently, and many scholars are still unsure whether it is worth publishing the

results of their work in digital form. Analyses of the innovative potential of digital scholarly editions can help editorial teams decide on the medium for the research effects presentation.

Using the example of selected digital correspondence editions, which are one of the most numerous groups of projects of this type, I intend to trace how the TEI standard affects areas (Maryl et al 2021) such as edition authors and recipients, data and software, and dissemination and evaluation.

Among the editions analyzed were such as *Maria Dabrowska - Anna Kowalska. Letters 1940-1965* published on the TEI Panorama platform or selected digital correspondences published on the TEI Publisher Registry such as Karl Barth Gesamtausgabe.

References

Maria Dabrowska - Anna Kowalska. Letters 1940-1965, <https://tei.nplp.pl/categories/korespondencja-marii-dabrowskiej-i-anny-kowalskiej>

Maryl, Maciej, Marta Błaszozyńska, Agnieszka Szulińska, et al. 2021. 'OPERAS-P Deliverable D6.5: Report on the Future of Scholarly Writing in SSH'. <https://doi.org/10.5281/zenodo.4922512>

Maryl, Maciej, Magdalena Wnuk, Françoise Gouzi, and Tomasz Umerle. 2024. 'Guidelines on Publishing and Evaluating Innovative Outputs in Social Sciences and Humanities', November. <https://zenodo.org/records/14221728>
TEI Publisher Registry, <https://www.e-editiones.org/map/>

Starting from Showing, Moving toward Sharing

Broadening TEI's Reach

Akihiro Kameda (1);
Kohei Ishii (2);
Satsuki Inoue (2);
Naoki Kokaze (2)

(1) National Institutes for the
Humanities, Japan;
(2) Chiba University, Japan

Keywords

TEI, HTML, research data management

Abstract

This study proposes Tidy Text Principles and workflows for structuring textual data. This approach begins with a meaning-oriented representation ("showing") aligned with specific research objectives and proceeds to full structuring and sharing. Two datasets are presented as case studies.

The first involves political documents and parliamentary records, using lightweight TEI XML aligned with the Parla-CLARIN schema, supplemented by HTML and JSON for visualization and annotation. The second addresses Wittgenstein's *Tractatus Logico-Philosophicus*, where an index dataset was restructured in JSON, then converted to TEI, focusing on mathematical content and multilingual term relationships. In both cases, preprocessing—transforming unstructured or inconsistently

structured texts into tidy forms—posed the main challenge, whereas TEI conversion itself proved straightforward once a tidy intermediate format had been established.

“Tidy text” is defined not by strict data normalization, but as a pragmatic, research-driven organization: units of interest are clearly segmented, minimally nested, and addressable. While TEI is often used as a starting point in textual studies, we suggest that workflows beginning elsewhere—due to practical or technical needs—can still reach TEI as a meaningful endpoint, particularly when tidy structuring is applied. The principles proposed herein remain preliminary and are intended to invite feedback during the presentation, with the aim of refining them collaboratively according to diverse research practices.

Additionally, we reconsidered the five-star model of Open Data in humanities research, where full open licensing or RDF-based linking is often impractical. We propose a revised model separating the provenance, licensing, and external linkage from core textual structuring. Our principle aligns with Levels 2–4 of the Best Practices for TEI in Libraries, but with lower overhead due to focused, purpose-driven markup.

By grounding structuring in research needs, the Tidy Text approach aims to support incremental, but interoperable documentation pathways, thereby broadening the reach of TEI in the digital transformation of scholarly practice.

Streamlining TEI Workflows

Collaborative Editing with NormaTEI

Pietro Sichera (1);
Salvatore Cristofaro (1);
Christian D’Agata (2);
Angelo Mario Del Grosso (1);
Miryam Grasso (2);
Laura Mazzagufò (1);
Daria Spampinato (1)

(1) Consiglio Nazionale Delle
Ricerche - CNR, Italy;
(2) Università degli Studi
di Catania, Italy

Keywords

Cooperative Digital Scholarly Edition,
Harmonization, Encoding, Interoperability,
NormaTEI

Abstract

In digital philology, the production of scholarly editions is increasingly recognized as a collaborative endeavour, involving multi-disciplinary teams and both synchronous and asynchronous workflows. This shift affects not only the publication of individual critical texts but also the construction of large-scale digital archives. While adopting formal protocols like those of the TEI is essential, it alone cannot guarantee consistency and coherence of the outcomes. A robust system is needed to support the entire editorial process, facilitating both the harmonization and quantitative analysis of texts.

In the literary domain, projects such as PirandelloNazionale,³ VerismoDigitale,⁴ and COVerLeSS⁵ have leveraged NormaTEI to ensure a common model for representing texts, variants, and multiple levels of annotation. This approach has facilitated integration between editions, lexicographic resources, and archival materials.

- 1 <https://bellinicorrespondence.cnr.it/>
- 2 <http://epicum.istc.cnr.it/>
- 3 <https://www.pirandellonazionale.it/>
- 4 <https://verismodigitale.uniot.it/>
- 5 <https://coverless.cnr.it/>

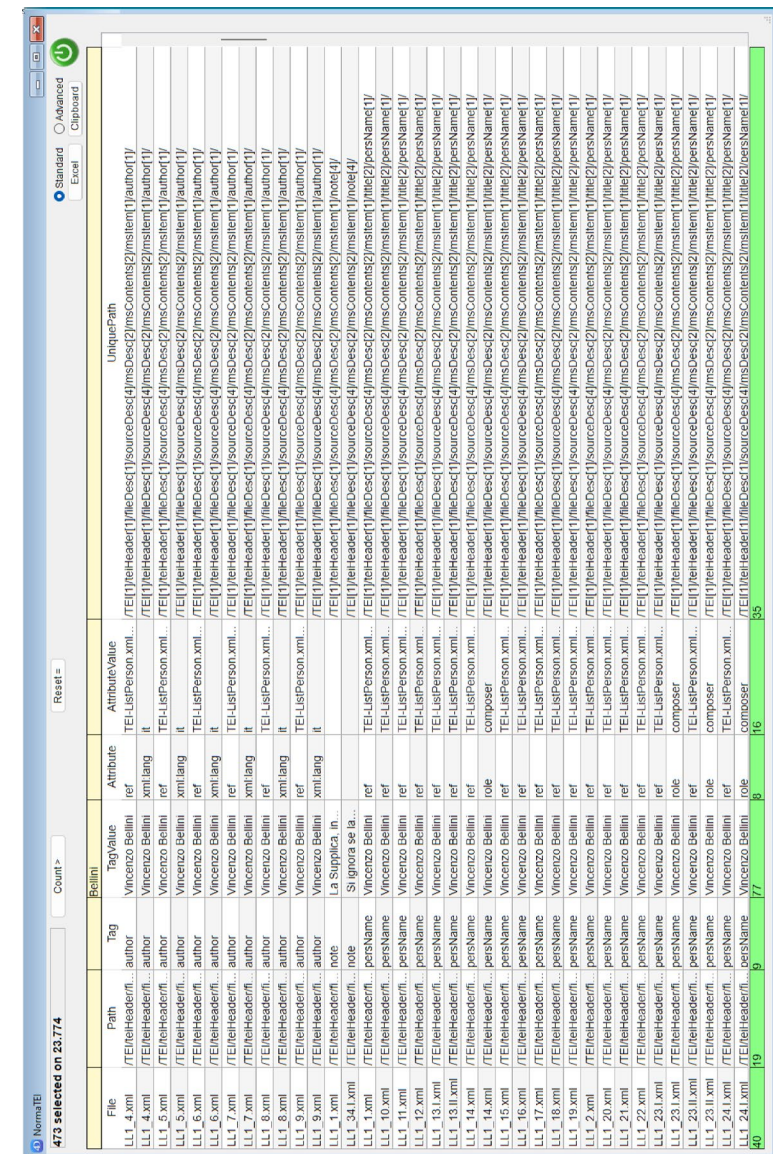


Fig. 1. Search interface of NormaTEI

A relevant use case occurred within the PAVES-e project,⁶ involving the “diplomatic” transcription of the poems in *Lavorare stanca* (Fig.2). The tool was used to identify over 70 different values for the @place attribute in <add> elements, compared to the 13 allowed by the project’s encoding model. NormaTEI made it possible to detect typographical errors, inconsistent formatting (e.g., CamelCase), unauthorized combinations, and even TEI-compliant values that did not conform to the project’s specific guidelines, though semantically valid.

NormaTEI has also been adopted for educational purposes. In degree classes in Digital Humanities at the University of Pisa and in Textual Studies for Digital Professions at the University of Catania, the tool has been used to introduce students to TEI encoding, supporting the learning of standard formats in textual markup and developing teamwork-related skills.

Moreover, NormaTEI has demonstrated versatility beyond TEI-encoded corpora. It was employed to review the encoding of Bellini’s autograph sketches in XML-MEI,⁷ addressing issues related to handwritten musical notation and the automated conversion from musical notation formats. Specifically, it enabled the correction of inconsistencies in the rendering of certain musical symbols, such as slurs, and the integration of authorial interventions into the already encoded text.

Although originally conceived as a tool for analyzing XML documents, NormaTEI has been evolving into an editor capable of batch-correcting structural and lexical discrepancies, including attribute values, identified during the analysis phase, while still refraining from direct intervention in the textual content of the elements. Finally, software requirements, licenses and documentation are available on GitHub.⁸

6 <https://digitalpavese.cnr.it/>
7 Music Encoding Initiative, <https://music-encoding.org/>.
8 <https://github.com/pierpaolosichera/NormaTEI.git>

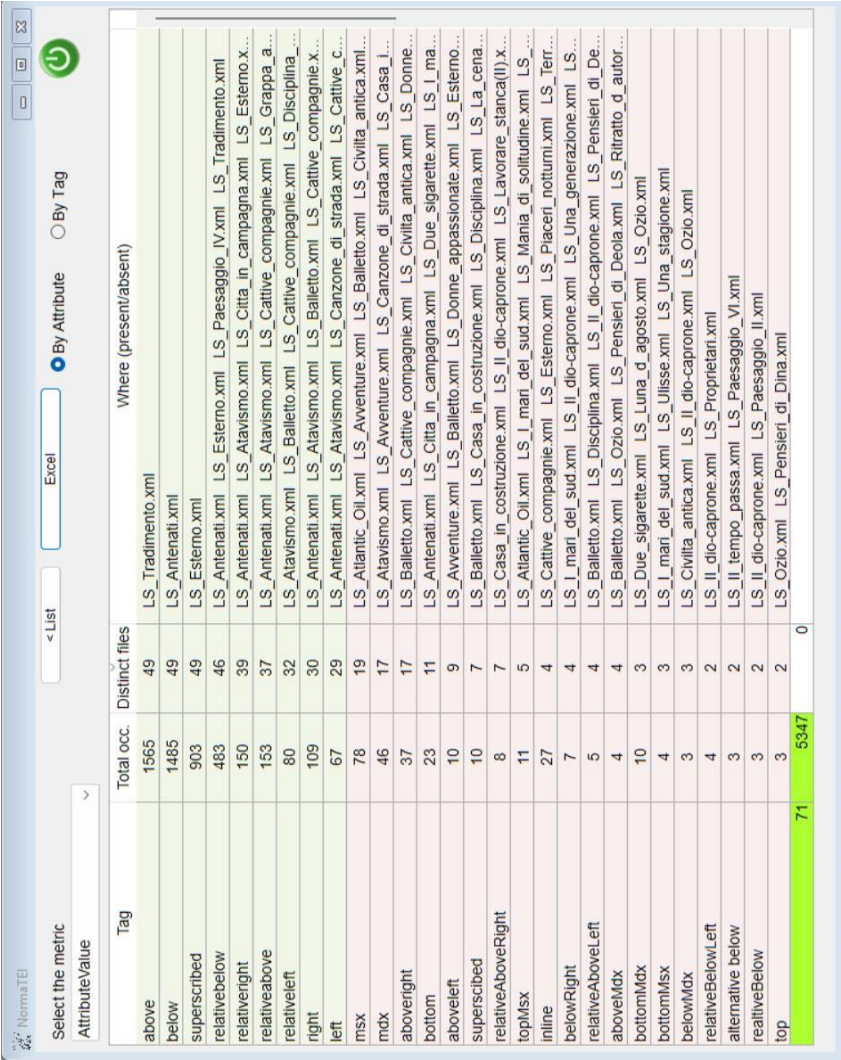


Fig. 2. Interface of NormaTEI, showing the first 28 occurrences of the different values of the @place attribute in the TEI <add> elements across the XML files encoding the poems in *Lavorare stanca*.

References

- Cristofaro, S., A. M. Del Grosso, L. Mazzagufu, P. Sichera, and D. Spampinato. 2025. "Implementing Collaborative Digital Scholarly Editions: Insights from *Bellini Digital Correspondence*." *IJIST* 9(1): 42–53. ISSN 2550-5114.
- Cummings, J. 2019. "Opening the Book: Data Models and Distractions in Digital Scholarly Editing." *International Journal of Digital Humanities* 1(2): 179–193. <https://doi.org/10.1007/s42803-019-00016-6>
- Cuculovic, M., F. Fondement, M. Devanne, J. Weber, and M. Hassenforder. 2022. "Semantios to the Rescue of Document-Based XML Diff: A JATS Case Study." *Software: Practice and Experience* 52(6): 1496–1516. <https://doi.org/10.1002/spe.3036>
- D'Agata, C., A. Di Silvestro, and A. Sichera. 2022. "Edizione critica, edizione digitale, Hyperedizione. "Il fu Mattia Pascal" come paradigma dell'edizione digitale dell'Opera Omnia di Luigi Pirandello." *Bollettino - Centro di studi filologici e linguistici siciliani* 33: 263–280. ISSN: 0577-277X.
- Mancinelli, T., and E. Pierazzo. 2020. *Che cos'è un'edizione scientifica digitale*. Roma: Carocci Editore.
- Robinson, P. 2017. "Some Principles for Making Collaborative Scholarly Editions in Digital Form." *Digital Humanities Quarterly* 11(2). <http://www.digitalhumanities.org/dhq/vol/11/2/000293/000293.html>
- Schmidt, D. 2014. "Towards an Interoperable Digital Scholarly Edition." *Journal of the Text Encoding Initiative* 7. <https://doi.org/10.4000/jtei.1072>
- Sichera, P. 2024. "NormaTEI: VO.6-beta (AIUCD2024)". *Zenodo*, 28 June 2024. <https://doi.org/10.5281/zenodo.12581646>

Structural Markup of the Database of Historical Polish Lexicons

The Case of *Forytarz języka polskiego* by Jan Ernesti and *Nowy dykcjonarz, to jest Mownik polsko-niemiecko-francuski* by Michał Abraham Troc

Ewa Rodek (1);
Łukasz Poznański (2)

(1) Institute of Polish Language Polish Academy of Sciences, Poland;
(2) Humanities Research Data Lab The John Paul II Catholic University of Lublin, Poland

Keywords

multilingual dictionaries, tagging lexicography, tagging a phrasebook, multi-search engine, multi-level headwords

Abstract

Based on two structurally different sources, we present marking the structure of historical dictionaries with Polish. In particular, we focus on the challenges involved in structurally marking up selected dictionaries using TEI markup, following the TEI Lex-O guidelines for lexicographic publications. The two works under analysis are *Forytarz języka polskiego* by Jan Ernesti (1674) and

Nowy dykajonarz, to jest mownik polsko-niemiecko-francuski, by Michał Abraham Troc (1764).

Historical dictionaries often lack consistent internal structures and are additionally affected by OCR errors, which significantly hinder automatic markup. As OCR typically does not preserve typeface or layout, punctuation becomes the primary cue for structuring.

Forytarz is a bilingual translation dictionary with a seemingly simple microstructure, though it presents several challenges. One issue involved identifying editorial corrections, which had been automatically marked during OCR with square brackets, despite serving different functions in the original. A more serious challenge was the inconsistent structure of German equivalents, which we deliberately chose not to normalize in order to preserve the author's intent. These equivalents are written identically to independent entries, using slashes – for example, *Blask / der Glantz der Sonnen / deß Feuers* (Fig. 1) versus *Pysk / die Raffe / Fresse* (Fig. 2). Furthermore, *Forytarz* contains a section of bilingual dialogues resembling a phrasebook or elementary language-learning manual. This required the inclusion of elements from the TEI *Drama* module (e.g., *<sp>*), alongside the lexicographic model, to appropriately encode speaker turns and dialogic structures.

By contrast, *Nowy dykajonarz* features a much more elaborate and regular microstructure. Nevertheless, punctuation again serves as the only reliable structural guide. Notable challenges include multi-level headwords and entries describing function words (see Fig. 3).

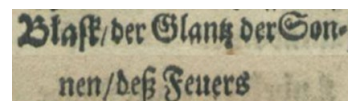


Fig. 1. Entry *Blask* in *Forytarz*... J. Ernesti, k. Avii



Fig. 2. Entry *Pysk* in *Forytarz*... J. Ernesti, k. Blij v

O. *interj.* 1) o, ach. 2) o; vor einem *Vocativo* in nachdrücklichen Redensarten. 1) o, ah, oh, hélas. 2) o; devant le *Vocatif* dans les expressions graves. § 2) o! bitogo i wzięto; o! świat tak płaci; o! już dość. 2) o nędzny świecie; o Boże moy.

O, *pr. Acc.* 1) um, wegen. pour, de. § o tę rzecz mię łaiano, bito; idźcie mi o maiętność; grać o co; idźcie o sławę; o groźz się powadzili; boję się o was wżysklich.

o co że; o zakład że. wir wollen wetten daß. voulez vous parier que; parions que.

2) auf; ohngefähr auf; um. environ, sur, à peu près; près de; vers. § o mieściac nad spodziewanie uchybił; o pięć fet ich było.

3) an. contre. § o kość zęb złamał; o rog koło się złamało; o skałę się okręt rozbił; o ziemię, o mur kim uderzyć; o głowę iego stłukł garniec.

* * *

In folgenden registret o den Localem. dans les sens suivans le Local suit après cette préposition o.

4) von, von einer Materie. sur, de quelque sujet. § o tey materyi mowił, piśał; o skromności życia rzecz miał; ia o cebuli, ty o czonku; prov. o tym gadka.

zle o tobie. es siehet schlecht um dich; es siehet schlecht mit dir aus. il y a du danger pour vous; vous êtes mal dans vos affaires.

5) um, in einer gefestten Zeit. à, dans un tems déterminé. § o tym czasie do domu nie chodzą; o piątey przyyde; o świtanu zaprzęgaj.

6) bey. au; à. § siedzi w więzieniu o chlebie i wodzie; o swoiemy strawie robi, służy.

7) mit. à; avec. § woz był o dwóch, o trzech kółach; o czterech koniach; pałac o tysiącu okien.

człowiek o dużym nosie. der Mensch mit der langen Nase. l'homme au grand nez.

8) bey. avec, en se soutenant, en s'appuyant sur qu. ch. § o kuli, o łafce chodzi.

* * *

o, pr. inf. dieses Vorwörtchen vor einem Verbo hat folgende Bedeutungen. cette préposition inséparable a les sens suivans devant un verbe.

1) um, in die Runde um. en, em, au tour. okręcić. umbreihen. entortiller. okrążyć. umzingeln, umgeben. entourer, environner.

2) ab, herunter, weg. de, en étant. odzierać. abreißen. détacher, dépouiller.

Ob, pr. inf. diese Präposition findet man vor vielen Verbis an statt O. on trouve cette préposition verbale souvent au lieu de la préposition O.

obłapić, obiać. umfassen, umarmen. embrasser.

Oba, Obadwa. beyde, alle beyde. deux, tous deux, tous les deux, tous deux ensemble.

Fig. 3. Entries describing the conjunction and preposition *O* in *Nowy dykajonarz*... M.A. Troc, col. 1081-1082

Structures and Tools for the Representation and Visualization of Knowledge Contained in Electronic Editions/Textual Resources

Iwona
Grabska-Gradzińska

Jagiellonian University, Poland

Keywords

computational ontologies, reasoning,
graph structures

Abstract

Electronic editions serve not only as reading resources but also as reservoirs of knowledge about the world – knowledge that is well-structured and encoded using standards that enable further processing and computational analysis. The TEI standard, based on XML, ensures a hierarchical and unambiguous data structure, which allows for the precise modelling of internal textual relationships. Such a structure guarantees organizational transparency of the text and facilitates its searchability, processability, and analysis using digital tools. The high level of encoding granularity further promotes editorial transparency and supports the comparison of multiple versions of a given text.

However, not all knowledge-related relationships embedded in the corpus of source texts are readily discernible to the

reader. Therefore, it is valuable to enhance textual perception using tools grounded in formal data representation. Treating the digital edition as a knowledge base accessible via an API allows external systems to interact with the data and perform reasoning based on the encoded content. Formal knowledge representation structures enable the imposition of an additional conceptual layer over the edition's architecture – one that organizes information in the form of concepts and relations and supports the application of computational ontologies and graph-based knowledge representation systems.

Ontologies provide formal definitions of concepts and the relationships among them. In the context of TEI-based textual editions, they can be employed to unambiguously define elements such as types of textual variants, person roles, or historical and cultural contexts. When combined with rule-based systems or inference engines, ontologies support the derivation of new information from existing datasets. Ultimately, this enables the construction of shared conceptual models, allowing for the comparison, integration, and interoperability of diverse digital editions.

A crucial role in knowledge processing is played by graph structures, as they effectively model relationships among concepts, entities, and events. Graph models, which are used to represent data, while enhanced by ontologies form knowledge graphs. In the context of knowledge representation and analysis – both within the digital humanities and in domains such as artificial intelligence and semantic systems – knowledge graphs offer a flexible and expressive model that supports both storage and inferencing.

The present work aims to demonstrate methods and strategies for representing knowledge derived from electronic editions developed within the Jagiellonian Digital Platform.

TEI Processing Model

Helena Bermúdez Sabel (2);
Magdalena Turska (1)

(1) e-editiones, Poland;
(2) JinnTec

Keywords

TEI Processing Model, TEI Publisher, XPath

Abstract

Crossing the divide between richly encoded XML sources and tangible, published digital edition has always been a weak spot for TEI community. The TEI Simple project aimed to bridge that gap with TEI Processing Model idea. Soon the TEI Guidelines were extended with the new `model` element and other provisions to represent and document editorial decisions about intended transformations. Since then, many editions worldwide embraced the approach and applied it with success in published digital scholarly editions and other academic publications – not only on the web, but also as e-books and printed publications.

Using the TEI Processing Model, customising the appearance of the text is all done in TEI ODD by mapping each TEI element to a limited set of well-defined behaviour functions, e.g. “paragraph” or “heading”. The TEI Processing Model specification includes a standard mapping, which can be tweaked by overwriting selected elements. Rendition styles are transparently translated into different output media types like HTML, XSL-FO,

LaTeX, or ePUB. This approach easily saves thousands of lines of code for media specific stylesheets.

Nevertheless, using the TEI Processing Model requires some familiarity with a number of technologies: the ability to formulate conditions in XPath, to express directives for presentation in CSS, to create snippets of HTML structures or to use custom web components for particular purposes, like creating a tooltip or embedding a facsimile viewer. These skills can be obtained through dedicated self-study, but the learning curve can be eased with a gentle in person introduction to the question. The proposed workshop intends to introduce the concepts of the TEI Processing Model and provide a tutorial on how to use TEI Publisher’s odd editor to create and elaborate a custom ODD.

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

The CMIFerator

A generalised pipeline for contributing to correspSearch

Julian Jarosch

Academy of Sciences and Literature
Mainz, Germany

Keywords

CMIF, correspSearch

Abstract

During the work on long-running digital scholarly editions of letters, it may become desirable to periodically generate and update Correspondence Metadata Interchange Format (CMIF) (TEI Correspondence SIG, 2025) files for use in correspSearch (Dumont et al., 2025). This may typically be based on a non-finalised working set of data, on data conforming to a superset of the CMIF schema (containing extraneous attributes or elements not covered by CMIF version 1.0), and/or on normalised data distributed over many individual letter files plus indices of persons, places and organisations (i.e. the structure required by ediarum (TELOTA, 2025).

From the observation of these requirements in a number of projects (Daugirdas and Kuczera, 2025; Dingel and Kohnle, 2023; Trautmann and Schrade, 2018; Wiese, 2025), the CMIFerator (Jarosch, 2025), a function library for eXist-db, was developed. It provides a set of functions for converting TEI letter files to CMIF XML. The design as a function library provides

some flexibility regarding which parts of the transformation are needed in the individual context. The full pipeline offered by the CMIFerator is:

- Update `<correspAction>` elements in individual letter files with the most up-to-date information from index files (regularised person names, person identifiers, regularised place names ...).
- Subset `<correspDesc>` elements in individual letter files for (strict) conformance to the CMIF standard.
- Wrap `<correspDesc>` elements in a CMIF template and fill in metadata.

The full pipeline is made accessible through convenient wrapper functions. At the same time, the three individual steps are available as separate functions and can be used individually or selectively if necessary.

The project-specific configuration is covered by a lean configuration file. At present, the CMIFerator intentionally is not conceived as a plug-and-play app, but instead example configurations, transformations and scripts are provided. These examples demonstrate how to use the function library in an XQuery script to either generate CMIF on the fly, or store it in the eXist database.

References

- Daugirdas, K., Kuczera, A. (Eds.), 2025. Zwischen Theologie, frühmoderner Naturwissenschaft und politischer Korrespondenz: Die sozinianischen Briefwechsel. Akademie der Wissenschaften und der Literatur, Mainz.
- Dingel, I., Kohnle, A. (Eds.), 2023. Flacius Briefwechsel. Digital. Akademie der Wissenschaften und der Literatur, Mainz.

Dumont, S., Grabsch, S., Müller-Laaackman, J., Sander, R., Sobkowski, S., 2025. correspSearch - Briefeditionen vernetzen.

Jarosch, J., 2025. CMIFerator.

TEI Correspondence SIG, 2025. Correspondence Metadata Interchange Format (CMIF).

TELOTA, 2025. ediarum - Digitale Editionen erstellen und publizieren.

Trautmann, M., Schrade, T. (Eds.), 2018. DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde. Akademie der Wissenschaften und der Literatur, Mainz.

Wiese, C. (Ed.), 2025. Buber-Korrespondenzen Digital: Das Dialogische Prinzip in Martin Bubers Gelehrten- und Intellektuellennetzwerken im 20. Jahrhundert. Akademie der Wissenschaften und der Literatur, Mainz.

The TEI Critical Apparatus at 30-something

Where do we go from here?

Hugh Cayless

Duke University, USA

Keywords

Critical Apparatus, Digital Critical Editions, TEI Guidelines

Abstract

The TEI's Critical Apparatus chapter seems to have first appeared in the TEI P3 Draft, published in print in 1994 (it is mentioned earlier in the P2 sources in the TEI Vault, but is not present there). It had then most of the features it has now, the app, lem, rdg, rdgGrp, wit, witDetail, the associated witStart/End and lacunaStart/End as well as witList (now listWit) and witness. Much of the prose and many of the examples in the chapter date back to that period. Substantial changes were made in 2015 and 2020 to permit the modeling of structural variation, tighten up the content model, and to improve the discussion of critical practices that the apparatus module may be used to implement. Despite these improvements, however, the Critical Apparatus chapter remains somewhat confusing and disjointed.

One example where the chapter takes a turn into obscurity is the section on "subvariation" (13.1.3) which introduces the

rdgGrp element and explains how it might be used to subdivide readings for the beginning of Chaucer's *Wife of Bath's Tale*. It should be noted that printed critical editions of Chaucer make little of this example, printing 'Experience' and noting the variant 'Experiment'. The reader is left wondering why they might commit to such elaborate encoding. A far better and more understandable use case for rdgGrp would be the encoding of variorum editions, where it can quite logically be used to collect the apparatus entries from previous editions of the work.

Part of the reason for the chapter's state is simply its age, and a reluctance on the part of the Guidelines' editors to commit to a thorough reworking, along with a certain deference to its original creators. Part, too, is the difficulty of the material. Further, the apparatus module is complicated by its attempt to do many things at once. It can model variation on a text, it can serve as a collation table, it can act as connective tissue between a set of varying sources, or it can serve to transcribe a previously printed apparatus. It features no less than three techniques for attaching the apparatus to the text, two of which are arguably obsolete in the P5 era, and none of which account adequately for the situation where an editor is presenting one apparatus with many texts. Moreover, as noted above, some of the examples recommend techniques for which it is hard to understand the desired outcome, being more theoretical than practical.

For all of these reasons, it is time for a re-consideration of the critical apparatus chapter, to prune some of the outdated recommendations and examples and to better account for recent work on digital critical editions. My talk will propose a new framework for organizing the chapter, centered around four basic use cases: single text critical editions, multitext critical editions, collation tables, and the encoding of existing print editions. It will suggest the chapter consider only two forms of apparatus connection: inline (i.e. parallel segmentation) and external.

Towards a model of transcultural encoding of ancient epigraphic sources

Michele Brunet (1);
Estelle Ingrand-Varenne (2);
Nicolas Souchon (5);
Nathalie Prévôt (3);
Vincent Razanajao (6);
Coline Ruiz-Darasse (3);
Blandine Nouvel (4);
Emmanuelle Morlock (1);
Bruno Baudoin (4);
Satre Stéphanie (4);
Rémi Bonnin (2);
Léontine Fortin (2);
Damien Strzelecki (2)

(1) Université Lumière Lyon 2, France, CNRS UMR 5189 HiSoMA, EquipEx Biblissima+;
(2) CNRS UMR 7302 CESCO, EquipEx Biblissima+;
(3) CNRS UMR 5607 Ausonius, Université Bordeaux Montaigne;
(4) CNRS, UMR 7299 Centre Camille Jullian, Aix Marseille Université;
(5) Institut français d'archéologie orientale;
(6) Centre d'études alexandrines, CNRS-IFAO

Keywords

Epigraphy, TEI, EpiDoc, Opentheso, Patrimonium Editor.

Abstract

This paper presents the work of the SA cluster, "TEI and Epigraphy", within the Biblissima+ consortium (Observatory of Written Cultures, from Clay to Print). In order to better capture the semantic interrelations between artefacts, images, and texts in epigraphic sources from Ancient Egypt, Greece, Rome, Gaul, and the Middle Ages, we have developed an integrated

digital environment. This environment combines a generic XML/TEI/EpiDoc template (GenEpiTemplate) with a structured multilingual thesaurus (GenEpiTheso), both accessible through a dedicated web-based editorial platform.

Biblissima+ provides a digital library and research infrastructure designed to support the study of the transmission of texts from Antiquity to the Modern era. As part of its broader scholarly objectives, the consortium initiated the integration of epigraphic data. An interdisciplinary team – comprising epigraphers working across diverse cultural and chronological contexts, together with specialists in Digital Humanities – was assembled to investigate unified methods of data representation. This cross-cultural approach contributes to refining the conceptual definition of the epigraphic document, emphasising its material, graphic, and iconographic dimensions.

The initial aim of the cluster was to establish a shared editorial template for its various projects, all of which rely upon the standards of the Text Encoding Initiative (TEI) and EpiDoc. Although a core model was rapidly identified, the first version revealed a number of limitations. Through comparative analysis of existing encoding practices, we developed an enriched and fully annotated template, extending EpiDoc with a broader range of TEI elements to accommodate a more extensive corpus of epigraphic material.

Concurrently, we developed a multilingual thesaurus, organised into seven thematic domains (artefact, material, text, language, script, production, and field of study). Each entry is accompanied by definitions, translations, and conceptual alignments, and is managed using the open-source platform Opentheso.

Both the encoding template and thesaurus are now accessible via the PETRAE portal and have been integrated into the PATRIMONIVM editor – an eXist-db application developed by the Ausonius Institute – facilitating the online editing of XML/TEI-EpiDoc encoded inscriptions. The platform is further connected to a Zotero library (GenEpiBiblio), which consolidates

essential bibliographic references across the principal domains of epigraphic research.

This presentation will introduce the project and its principal outcomes, and wishes to open a discussion with TEI specialists concerning the methodological and technical challenges encountered during its development.

Towards an Encoding Practice for Multilingual Textual Variation

Sandra Balck;
Fabian Etling

Freie Universität Berlin, Germany

Keywords

digital scholarly edition, multilingualism,
text comparison, textual criticism

Abstract

Multilingual text versions pose challenges for digital scholarly editions (DSE), particularly in the encoding and representation of textual variation. Hannah Arendt for example authored most of her works in English and German, carefully rewriting and editing the texts herself. The resulting versions should not be considered as mere translations or revisions. Arendt used the plurality of languages as a creative tool. In comparing the versions, seeming textual shifts create meaningful tensions. For a DSE, this imposes the task of enabling readers of Arendt's works to explore "these tensions as a significant and signifying *Between*." (Wild, 2024, p. 29)

The COMUTE project⁹ was initiated to analyse such "tensions" in multilingual writing. Funded by the German Research Foundation, it aims to algorithmically identify variant passages¹⁰

⁹ <https://www.comute-project.de/>

¹⁰ The COMUTE-Project is building on the existing monolingual collation tool LERA (Pöckelmann et al., 2022) and is extending it by adding a multilingual layer.

and, in collaboration with the Hannah Arendt Critical Edition¹¹ as one use case, explores how such findings can be evaluated, visualised and integrated into editorial workflows. In order to support data reuse, especially in DSE workflows, the collation results should be encoded according to the TEI guidelines.¹²

While it is feasible to use TEI markup for critical apparatuses for simpler cases, this is not straightforward when it comes to more complex cases, as phenomena of multilingual textual variation manifest themselves not only at the character level, but also in linguistic nuances and in content and context. This poster¹³ presents a TEI-based approach for encoding multilingual and non-parallel textual variation and integrating it into DSEs.

References

- Bleeker, E., Buitendijk, B., Haentjens, R. and Kulsdom, A. (2018) 'Including XML Markup in the Automated Collation of Literary Text', *XML Prague 2018 Conference Proceedings*, pp. 77-95. Available at: <https://archive.xmlprague.cz/2018/files/xml-prague-2018-proceedings.pdf>
- Cayless, H., Beshero-Bondar, E., Vigilanti, R. and Cummings, J. (2019) 'Document Modeling with the TEI Critical Apparatus', *TEI 2019: What is text, really? TEI and beyond*, University of Graz, Austria, 16-20 September. *Book of Abstracts*, pp. 168-170. Available at: https://gams.uni-graz.at/o:tei2019_bookofabstracts

¹¹ <https://hannah-arendt-edition.net/home>

¹² As discussed for monolingual variation, for example, by Bleeker et al. (2018) and Cayless et al. (2019) and for translation scenarios in the TEI mailing list (Various Authors, 2013).

¹³ DOI: <https://doi.org/10.5281/zenodo.15834844>

Pöckelmann, M., Medek, A., Ritter, J. and Molitor, P. (2023) 'LERA - An interactive platform for synoptical representations of multiple text witnesses', *Digital Scholarship in the Humanities*, 38(1), pp. 330-346. Available at: <https://doi.org/10.1093/lc/fqac021>

Various authors (2013) 'Translation as witness', *TEI (Text Encoding Initiative) public discussion list*, 6 November 2013 13:56:01 - 0500. Archive of thread available at: <https://lists.psu.edu/cgi-bin/wa?A2=1311&L=TEI-L&D=0&P=15199371>

Wild, T. (2024) 'Relational Reading: Hannah Arendt's The Human Condition and Vita activa, and The Plurality of Languages', *Angermion*, 17(1), pp. 139-158. Available at: <https://doi.org/10.1515/anger-2024-0005>

Voci dall'Inferno

a TEI-Based Digital Archive for finding Dante in Concentration Camp Testimonies

Elvira Mercatanti (1);
Marina Riccucci (2);
Angelo Mario Del
Grosso (1)

(1) ILC: CNR-Istituto di Linguistica
Computazionale "A. Zampolli",
Pisa, Italy;
(2) Università di Pisa, Italy

Keywords

XML-TEI, digital archives, eXist-db, XQuery,
web application

Abstract

Voci dall'Inferno is a digital humanities project that brings together a multidisciplinary team of scholars and Digital Humanities students. The project pursues two interconnected objectives: (a) the creation of the first digital corpus of non-literary testimonies of concentration camp survivors - mostly unpublished - encoded in XML-TEI, including both written and oral sources (Burnard, 2014); (b) the systematic identification and analysis, within the testimonies, of lexical items and allusions to Dante's *Divine Comedy*, particularly the *Inferno* (Calderini and Riccucci, 2020; Mattingly, n.d.).

The project is grounded in the hypothesis that Dante's lexicon provided survivors with an expressive framework to talk about the ineffable experience of the Lager, overcoming the limits of its own personal vocabulary. The frequent and meaningful recurrence of Dantean references - regardless of educational

background - suggests the *Commedia* acts as a shared cultural code (Calderini and Riccucci, 2020).

The ongoing digital corpus is accessible through a web application developed within the eXist-db environment using XQuery, which supports querying, annotation visualization and intertextual analysis (see a demo at: <https://tinyurl.com/mkoez3f8>). Encoding strictly follows profitably TEI Guidelines, employing modules for encoding oral data and modules for primary written sources as well as named entities and an extensive metadata recording. A parallel transcription approach was adopted: TEI *zone* elements for written sources and TEI *timeline* mechanisms for oral ones. The corpus is developed in line with FAIR principles (Del Grosso et al., 2024b).

The archive includes 23 testimonies from 18 witnesses (13 oral, 10 written), totalling 18 hours and 35 minutes of audio and 390 transcribed pages. Preliminary analysis reveals 61 references to the *Comedy*, highlighting how Dante's *Inferno* serves as a semantic bridge between literary memory and the survivors' needs (Fig. 1) (Del Grosso et al., 2024a).

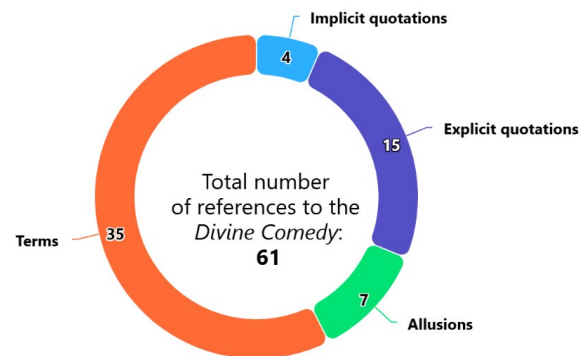


Figure 1: distribution of Dante's lexicon across the corpus: a total of 61 references, including 15 explicit quotations, 4 implicit quotations, 7 allusions, and 35 terms.

Our current work focuses on the automatic text recognition task, using tools like Whisper and eScriptorium, and the automatic extraction of Dante's quotations (Fig. 2), using Sentence-BERT and VectorializedDB.

Voci dall'Inferno Verse Similarity Search

Enter your search query:

E caddi come corpo morto cade.

Select canto/i

Choose an option

Similarity threshold: 0.50 Number of results: 5

0.00 1.00 1 10

Search

Found 5 similar verses:

V (Similarity: 1.00)

E caddi come corpo morto cade.

III (Similarity: 0.65)

e caddi come l'uom cui sonno piglia.

Figure 2: *Voci dall'Inferno Verse Similarity Search*: a sentence from a testimony is used as input, and the output returns the verses from Dante's *Divine Comedy* that are most semantically similar.

This paper will present the TEI encoding strategies adopted as well as the broader methodological implications of the project.

References

- Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. Marseille: OpenEdition Press. <https://doi.org/10.4000/books.oep.426>
- Calamai, S., Beeken, J., Scagliola, S., Broekhuizen, M., Draxler, C., Van Hessen, A. and Van den Heuvel, H. (2021). 'Voices from Ravensbrück: Towards the creation of an oral and multilingual resource family'. In: *Proceedings of CLARIN Annual Conference 2021*, pp.16-19.
- Calderini, S. and Riccucci, M. (2020). 'L'ineffabilità della nefandezza: Dante "per dire" il Lager: un sondaggio preliminare nelle testimonianze non letterarie'. *Italianistica*, 49, pp.213-228. <https://doi.org/10.19272/202001301011>
- Del Grosso, A.M., Riccucci, M. and Mercatanti, E. (2024a). 'Voci dall'Inferno: Dante per dire il Lager - Digitalizzare e studiare le testimonianze'. In: *ME.TE. Digitali: Mediterraneo in rete tra testi e contesti*. Catania: AIUCD.
- Del Grosso, A.M., Riccucci, M. and Mercatanti, E. (2024b). 'The impact of digital editing on the study of Holocaust survivors' testimonies in the context of Voci dall'Inferno project'. In: Anuradha, I., Wynne, M., Frontini, F. and Plum, A. (eds.) *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*. Torino, Italia: ELRA and ICOL, pp.1-9. <https://aclanthology.org/2024.htres-1.1/>
- Levi, P. (1997). *Conversazioni e interviste (1963-1987)*. Belpoliti, M. (ed.). Torino: Einaudi.
- Mattingly, W. (n.d.). *wjbmattlingly/vulgata-spacy*. [Online] Available at: <https://github.com/wjbmattlingly/vulgata-spacy> (Accessed 26 June 2025).
- Streamlit (n.d.). *Latin Vulgate App*. [Online] Available at: <https://latin-vulgate.streamlit.app/> (Accessed 26 June 2025).

What do we Need to Make Documents from Texts?

Georg Vogeler

University of Graz, Austria

Keywords

documentality, authentication, diplomatics

Abstract

The term *document* appears frequently in the TEI Guidelines, generally with three meanings: (1) any textual object (e.g., in `<docTitle>`), (2) a computer file containing user information (e.g., the `<TEI>` element), and (3) a material object bearing conventional symbols (e.g., in `<sourceDoc>`). This usage reflects general language, as captured by dictionaries like Merriam-Webster. Within the TEI community, and especially among digital scholarly editors, a narrower focus has prevailed, centered on the third definition and practices associated with "documentary editing."

A broader perspective is offered by Maurizio Ferraris, who proposes a theory of *documentality*—seeing documents as fundamental social artefacts that define social reality (Ferraris 2009, 2011, Ferraris & Torrenco 2014). This view aligns with the historical discipline of diplomatics, which studies methods developed to authenticate and manage social relationships through documents. Diplomats has thus become both a field of historical forensics (*discrimen veri ac falsi*) and a cultural history of documents. Luciana Duranti (1989, 1998) has continued this tradition into modern archival science.

Central to diplomatics is the *authenticity* of documents: establishing verifiable links between a text and the individuals creating the social relationship it records. Historically, this was achieved through public acts, reliance on trusted officials like notaries, the development of stylistic norms (e.g., in papal chancery), and material signs like signatures and seals. Today, digital methods such as hashes and cryptographic signatures fulfil similar roles. Sean Winslow (2020) has proposed how TEI encoding could better capture these features. Meanwhile, the *Vocabulaire international de la diplomatie* (VID, Càrcel Ortí 1997)¹⁴ provides an established terminology for describing these aspects, showing their relevance across cultures (Vogeler 2018).

However, the TEI Guidelines currently lack an ontological model for representing the documentary status of texts in the diplomatic sense. Although diplomatic documents have been encoded with TEI, often semantically reduced (e.g., Isasi Martínez & Spence 2014; MRSH Caen¹⁵), or by integrating the VID (as in Diple¹⁶ or Vujosevic & Vogeler 2025), these efforts remain marginal. The Charters Encoding Initiative (CEI)¹⁷ represents a more formal attempt but is based on TEI P4. Yet these initiatives expose a deeper issue: the TEI's conceptual model does not recognize the full social and forensic nature of documents.

Diplomatic concepts like the *authenticum* or copy chains crucial to establishing authenticity are poorly represented. TEI's `<signed>` and `<seal>` elements encode textual and physical features without connecting them to their authenticating functions. `<filiation>` remains purely relational.

In 2022, Sean Winslow submitted a formal proposal to better accommodate diplomatic concepts in TEI.¹⁸ While the commu-

¹⁴ <https://www.cei.lmu.de/VID/>

¹⁵ <https://mrsh.unicaen.fr/editer-des-sources-diplomatiques/>

¹⁶ <http://developpements.enc.sorbonne.fr/diple/schema/acte/>

¹⁷ <https://www.cei.lmu.de>

¹⁸ <https://github.com/TEIC/TEI/issues/2376>

nity's cautious response—suggesting, for instance, the use of customization profiles—is understandable, I argue that this reluctance reflects a deeper neglect of essential textual concepts.

In my presentation, I will advocate for rather small extensions to the *teiHeader* to represent the documentary and forensic dimensions of texts: a new `<authDesc>` as property of the TEI document or an individual textual witness. `<authDesc>` can include descriptors of means of authentication like seals, stamps, signatures, md5 hashes and similar or a prose reflection on the authenticity status. A controlled vocabulary for the core authenticity status (original, forged, suspicious, unclear) supports machine readability. An element focussing on the roles in text creation and processing that support its documentary status as well as the role in the social act (`<legalActor>`) support this and can be detailed by referencing established external sources like mentioned *Vocabulaire Internationale de la Diplomatie*.

I invite the TEI community to discuss whether we should simply accept the status quo—or whether we can, and should, open the TEI Guidelines to communities focused on documentality: digital archivists, legal scholars, and—certainly—students of historical documentary cultures.

References

- Càrcel Ortí, Maria Milagros. 1997. *Vocabulaire International de La Diplomatie*. 2nd ed. Col·lecció Oberta. Valencia: Univ. de València.
- Duranti, Luciana. 1989. 'Diplomatics : New Uses for an Old Science'. In *Archivaria* 28, 28:7-27. Ottawa. https://www.academia.edu/5404671/Diplomatics_New_Uses_for_an_Old_Science.
- . 1998. *Diplomatics: New Uses for an Old Science*. Lanham: Scarecrow Press.
- Ferraris, Maurizio. 2009. *Documentalità. Perché è Necessario Lasciar Tracce* (Eng. Tr. by R. Davies, *Documentality: Why It Is*

- Necessary to Leave Traces*, New York, Fordham University Press, 2012). Roma-Bari: Laterza.
- . 2011. 'Social Ontology and Documentality'. In *Approaches to Legal Ontologies*, edited by Giovanni Sartor, Pompeu Casanovas, Mariangela Biasiotti, and Meritxell Fernández-Barrera, 83–97. Law, Governance and Technology Series 1. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-94-007-0120-5_5
- Ferraris, Maurizio, and Giuliano Torrenzo. 2014. 'Documentality: A Theory of Social Reality'. *Rivista Di Estetica* 57:11–27. <https://doi.org/10.4000/estetica.629>
- Isasi Martínez, Carmen, and Paul Spence, eds. 2014. 'Guía Para Editar Textos CHARTA Según El Estándar TEI: Una Propuesta.' <https://www.redcharta.es/investigacion/>
- Vogeler, Georg. 2013. 'Von der Terminologie zur Ontologie : Das "Vocabulaire international de la diplomatique" als Ressource des Semantic Web'. *Francia* 40:281–97.
- Vogeler, Georg. 2018. 'Digital Diplomatics: The Evolution of a European Tradition or a Generic Concept?' In *Studies in Historical Documents from Nepal and India*, edited by Simon Cubelic, Astrid Zotter, and Axel Michaels, 1:85–109. Documenta Nepalica. Heidelberg: Heidelberg University Publishing. <https://doi.org/10.17885/heiup.331.c4129>
- Vujošević, Žarko, and Georg Vogeler. 2025. 'Encoding Medieval Charters in Preparation for a Diplomatic Semantic Web. Combining the TEI with the Vocabulaire International de Diplomatie'. Zenodo. <https://doi.org/10.5281/zenodo.14696237>
- Winslow, Sean M. 2020. 'Authenticating Features in the TEI'. *Journal of the Text Encoding Initiative*, no. Issue 13 (May). <https://doi.org/10.4000/jtei.3608>

Who Knows What a Revision Is?

Towards a Shared Vocabulary of Textual Variation

Elli Bleeker (1);
Beatrice Nava (2)

(1) The Huygens Institute,
Netherlands;
(2) University of Vienna, Austria

Keywords

textual variation, collation, visualization,
ontology, textual scholarship

Abstract

This paper presents preliminary results from an initiative introduced at the international workshop Seeing the Difference: Visualizing Textual Variation (Leiden, April 2024), aimed at aligning scholarly vocabularies used to describe textual variation. While the idea of a mapping of textual scholarship terminology is not new (Dillen 2020; Macé and Roelli, eds.), the need remains. For example, the understanding of "revision" differs across historical periods and material contexts, despite using the same TEI-encoding. The great advantage of the TEI is that it offers textual scholars flexibility in recording and categorizing variation, however, this lack of standardization is also a core problem for the development of tools and visualisations of textual change, as it prevents interoperability, replicability and evaluation of results across textual traditions. Therefore, consensus on vocabulary and specifications for functional

visualizations is crucial to developing broadly useful tools, facilitating data exchange, and building larger corpora for textual variation analysis.

The workshop's resulting working group addresses this problem by uniting scholars from different periods and specialities. This paper reports on the ongoing work of reaching a shared agreement: we document usage patterns, map relationships between terms, and develop definitions acknowledging contextual differences. Starting with use cases collecting examples of differently interpreted TEI elements, we develop a preliminary shared vocabulary, formalize it in an ontology, building upon existing research (Giovannetti; Christen and Spadini 2029), and map it to TEI elements. Specifically, the paper presents 5 key concepts understood differently across traditions, draft definitions acknowledging these differences, and a knowledge graph showing relationships between concepts. The preliminary results will be presented at the conference as a call for participation in further developing the ontology for textual variation.

References

- Christen, A., & Spadini, E. (2019). "Modeling genetic networks. Gustave Roud's œuvre, from diary to poetry collections". *Umanistica Digitale*, 3 (7). <https://doi.org/10.6092/issn.2532-8816/9063>;
- Dillen, W. (2020). "A Lexicon of Scholarly Editing" (v.O.4.1). Zenodo. <https://doi.org/10.5281/zenodo.4008439>;
- Giovannetti, F. Critical Apparatus Ontology (CAO). <https://w3id.org/cao>;
- Macé, C. and P. Roelli, eds. "Parvum lexicon stemmatologicum". <https://wiki.helsinki.fi/xwiki/bin/view/stemmatology/Parvum%20lexicon%20stemmatologicum/>, last modified on 2024/02/13.

Program Committee

- Syd Baumann**, Northeastern University, USA
Helena Bermúdez Sabel, JinnTec, Spain
Elisa Beshero-Bondar, Penn State Erie, USA
Peter Boot, Huygens Institute, The Netherlands
Hugh Cayless, Duke University, USA
Constance Crompton, University of Ottawa, Canada
James Cummings, Newcastle University, UK
Maciej Eder, Polish Academy of Sciences, Poland
Maria Fronczak, Polish Academy of Sciences, Poland
Martin Holmes, University of Victoria, Canada
Diane Jakacki, Bucknell University, USA
Katarzyna Anna Kapitan, University of Oxford, UK
Magdalena Komorowska, Jagiellonian University, Poland
Wolfgang Meier, e-editiones / TEI Publisher, Germany
Maciej Mikula, Jagiellonian University, Poland
Gimena del Rio Riande, CONICET, Argentina
Martina Scholger, University of Graz, Austria
Sabine Seifert, University of Potsdam, Germany
Anna Skolimowska, University of Warsaw, Poland
Bogumił Szady, Polish Academy of Sciences, Poland
Agnieszka Szulińska, Polish Academy of Sciences, Poland
Grażyna Ślusarczyk, Jagiellonian University, Cracow
Magdalena Turska, e-editiones / TEI Publisher, Poland (chair)
Joey Takeda, Simon Fraser University, Canada
Raff Vigilanti, University of Maryland, USA
Aneta Wysztygiel, University of Warsaw, Poland

Local Organizers

Iwona Grabska-Gradzińska

Joanna Hałaczekiewicz

Magdalena Komorowska

Agata Kwaśnicka-Janowicz

Jan Rybicki

Aleksandra Rykowska

Local Advisory Board

Agata Holobut

Maciej Mikula

Jeremi Ochab

Wojciech Słomczyński

Magdalena Turska

Magdalena Wójcik