Text Corpus in Collaboration

A Balance Between Customized and Standardized Approach

Summary. The information infrastructure subproject (INF) utilizes a TEI-XML based schema to integrate textual data from multiple subprojects, each rooted in different linguistic, temporal, and disciplinary contexts. Rather than replicating the original sources—which span scholarly editions, linguistic corpora, and educational materials in formats like plain text, HTML, XML, and tabular data such as CONLLU—the primary goal is to enable digital annotations that support metaphor analysis. While existing annotation tools such as Recogito, or INCEpTION provide useful platforms, they often fall short when dealing with structurally rich, multilingual, or historically complex texts. Essential features such as sectioning, special character encoding, complex text layouts (e.g., ruby annotations), or multiple tokenization layers—as required for languages like Sanskrit—are difficult to model using linear text representations. To meet these demands, the INF project developed an encoding approach grounded in TEI P5 and bridges the gap between standardization and customization. This model balances simplicity for annotation tasks with the flexibility to capture diverse textual phenomena. It introduces a shared structural backbone across all documents, facilitating consistent navigation and annotation, while remaining adaptable to specific project needs. It supports a reusable digital corpus, enabling a scalable infrastructure for collaborative research in the Digital Humanities.

Keywords: data reuse, text encoding initiative, annotation

1 Introduction

When creating corpora in the (digital) humanities, FAIR data principles are now increasingly established, making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). But many existing text collections that are highly relevant for research in the humanities—ranging from online databases and community archives to historical print editions—have not been created with the same focus on reuse. As a result, we are confronted with the task of reusing corpora that were never designed to be reused. This observation was the starting point for the research data management in the Collaborative Research Center (CRC) 1475, “Metaphors of Religion,”[1] which unites 14 subprojects that investigate metaphorical language across religious traditions,

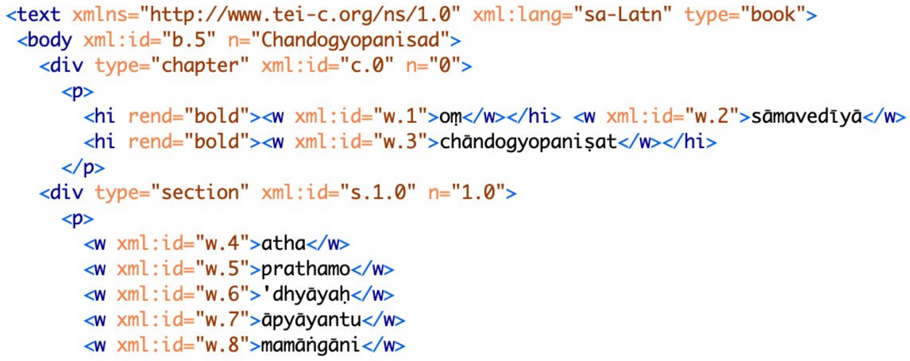
[1] Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958.

with sources ranging from ancient Sumerian and Biblical Hebrew to contemporary German and English, spanning from 3000 BCE to today. While incorporating the diverse source data into our infrastructure, our goal is not to re-create the original resources like scholarly editions or linguistic databases. The primary reason we incorporate the data into our own infrastructure is to enable digital annotations as a means to add an interpretative layer—in our case concerning metaphor analysis. Many annotation tools like CATMA (Gius et al. [2008] 2025) or Recogito (Barker et al. [2016] 2019) operate on plain text. More linguistically oriented annotation environments like INCepTION (Klie et al. 2018) work on streams of tokens that can carry additional information and are often represented in tabular form. Such approaches pose challenges for the kind of data we are working with:

- Structural information like sections or discourse structure is lost.
- Not all signs can be represented in a plain text version, like characters not (yet) in Unicode, but also pictorial elements like custom emoji in online forums.
- Complex text layout might be essential for understanding the text, like ruby annotations, which cannot be represented faithfully in a strictly linear text model.
- Tokenization is non-trivial for many historical and non-Western languages, meaning that the text model needs to be able to represent multiple tokenization layers for a single text.

Thus, we needed to create a text model that is simple enough to support the main goal of text annotation, but also flexible enough to capture and display more complex textual phenomena that cannot be represented in a simple plain text model.

Our unified text model is based on TEI P5 (TEI Consortium 2025) and preserves structural information like sections, as well as tokens that serve as annotation targets (shown in figure 1). The next step involves adapting diverse source texts into this framework—a process that involves both generalized and specialized approaches.



```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:lang="sa-Latn" type="book">
 <body xml:id="b.5" n="Chandogyopanisad">
   <div type="chapter" xml:id="c.0" n="0">
     <p>
       <hi rend="bold"><w xml:id="w.1">oṃ</w></hi> <w xml:id="w.2">sāmavedīyā</w>
       <hi rend="bold"><w xml:id="w.3">chāndogyopaniṣat</w></hi>
     </p>
   <div type="section" xml:id="s.1.0" n="1.0">
     <p>
       <w xml:id="w.4">atha</w>
       <w xml:id="w.5">prathamo</w>
       <w xml:id="w.6">'dhyāyaḥ</w>
       <w xml:id="w.7">āpyāyantu</w>
       <w xml:id="w.8">mamāṅgāni</w>
```

Figure 1. Structure of a common TEI text body

2 Generalized Approaches

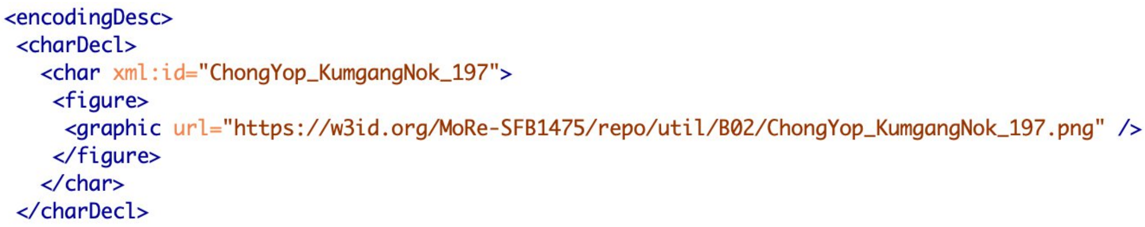
2.1 Handling Incomplete or Unreadable Texts

Historical texts frequently include missing or unreadable characters, particularly in damaged manuscripts or fragmented clay tablets. To address these challenges, we follow the EpiDoc framework (Elliot et al. [2017] 2025). While various disciplines use different methods for documenting editorial interventions, the Leiden system has become a widely accepted standard in philology and related fields. The EpiDoc Cheatsheet[2] translates these conventions into consistent markup, which is essential for the CRC's comparative work. With the help of CSS styling, this markup can be flexibly rendered to suit the specific needs of individual subprojects.

2.2 Missing UTF-8 Characters

Some characters in historical Chinese and Korean texts are not available in Unicode and pose significant challenges for representation. We represent them as custom glyphs using the *gaiji* TEI module images (shown in figure 2). We use the same method for custom emoji in forum data (Reimann et al. 2024), highlighting our reusable encoding strategy that can accommodate both rare historical characters and contemporary digital symbols.

[2] https://svn.code.sf.net/p/epidoc/code/trunk/guidelines/msword/cheatsheet.pdf



```
<encodingDesc>
 <charDecl>
   <char xml:id="ChongYop_KumgangNok_197">
    <figure>
     <graphic url="https://w3id.org/MoRe-SFB1475/repo/util/B02/ChongYop_KumgangNok_197.png" />
    </figure>
   </char>
 </charDecl>
```

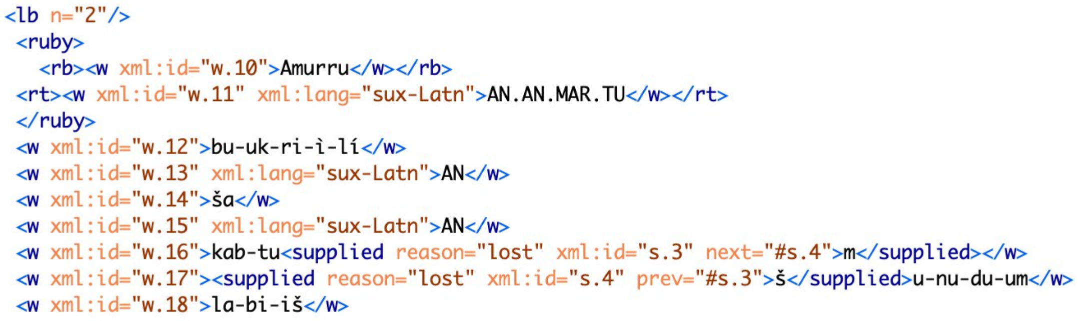
Figure 2. Referencing missing UTF-8 characters with images

3 Specialized Approach

Standardizing these multilingual texts into a generic format has its limitations, since some languages/phenomena simply require special treatment.

3.1 Ruby Annotations

Ruby annotations are small, supplementary text attached to the main text to provide pronunciation, glosses, or meanings, often used in East Asian languages. In our CRC, we use the TEI ruby module for Akkadian language texts which contain Sumerian pronunciations/versions of some words (shown in figure 3).



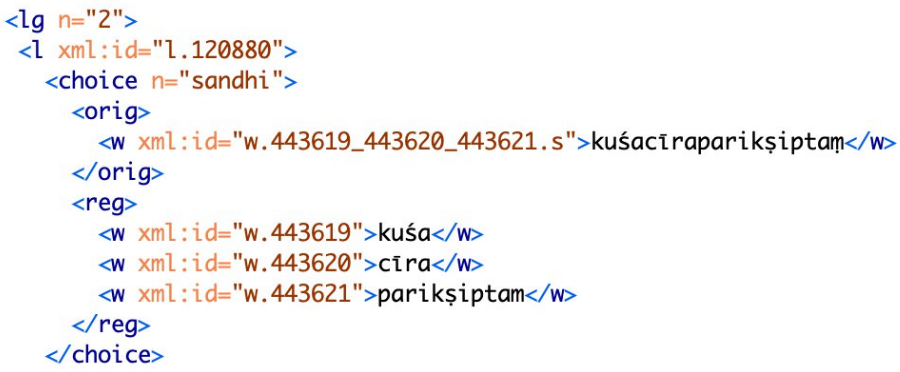
```
<lb n="2"/>
 <ruby>
   <rb><w xml:id="w.10">Amurru</w></rb>
 <rt><w xml:id="w.11" xml:lang="sux-Latn">AN.AN.MAR.TU</w></rt>
 </ruby>
 <w xml:id="w.12">bu-uk-ri-ì-lí</w>
 <w xml:id="w.13" xml:lang="sux-Latn">AN</w>
 <w xml:id="w.14">ša</w>
 <w xml:id="w.15" xml:lang="sux-Latn">AN</w>
 <w xml:id="w.16">kab-tu<supplied reason="lost" xml:id="s.3" next="#s.4">m</supplied></w>
 <w xml:id="w.17"><supplied reason="lost" xml:id="s.4" prev="#s.3">š</supplied>u-nu-du-um</w>
 <w xml:id="w.18">la-bi-iš</w>
```

Figure 3. Ruby annotation

3.2 Multiple Representations of Words

For texts available in classical Sanskrit, we have words available at both sandhi (compound) and unsandhi (or vigraha, the process of breaking these compounds into their original forms for analysis) level.

It poses a challenge due to sandhi rules (Elwert et al. 2015), where words combine, altering their original forms. In our CRC, we have stored both versions in a single XML file and still refer to each version as needed (shown in figure 4). We use the same method for showing vocalized and unvocalized forms of Hebrew text.



```
<lg n="2">
 <l xml:id="l.120880">
   <choice n="sandhi">
     <orig>
       <w xml:id="w.443619_443620_443621.s">kuśacīraparikṣiptaṃ</w>
     </orig>
     <reg>
       <w xml:id="w.443619">kuśa</w>
       <w xml:id="w.443620">cīra</w>
       <w xml:id="w.443621">parikṣiptam</w>
     </reg>
   </choice>
```

Figure 4. Sandhi-unsandhi words

To reconcile language-specific requirements with broad interoperability, we design a standardized encoding approach with core elements for downstream processing, while supporting extensible structures to capture unique linguistic features. This ensures precise, reusable representations across languages, enhancing the system's robustness and versatility.

Bibliography

- Barker, Elton, Leif Isaksen, Rebecca Kahn, Rainer Simon, and Valeria Vitale. (2016) 2019. *Recogito 2*. Scala. January 19; Pelagios Network, released. https://github.com/pelagios/recogito2.
- Elliot, Tom, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, and Scott Vanderbilt. (2017) 2025. *EpiDoc Guidelines: Ancient Documents in TEI XML*. V. 9. released. https://epidoc.stoa.org/gl/latest/.
- Elwert, Frederik, Sven Sellmer, Sven Wortmann, Manuel Pachurka, Jürgen Knauth, and David Alfter. 2015. "Toiling with the Pāli Canon." In *Proceedings of the Workshop on Corpus-Based Research in the Humanities*, edited by Francesco Mambrini, Marco Passarotti, and Caroline Sporleder. https://doi.org/10.15496/publikation-52722.

- Gius, Evelyn, Jan Christoph Meister, Marco Petris, et al. (2008) 2025. *CATMA*. V. 7.2.0. Zenodo, released March 10. https://doi.org/10.5281/ZENODO.1470118.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckhart de Castillo, and Iryna Gurevych. 2018. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.” *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9.
- Reimann, Sebastian, Lina Rodenhausen, Frederik Elwert, and Tatjana Scheffler. 2024. “By a Thread: Encoding Online Forum Data in TEI.” *Journal of the Text Encoding Initiative*, no. Issue 17 (April): Issue 17. https://doi.org/10.4000/1209k.
- TEI Consortium. 2025. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. V. 4.9.0. TEI Consortium, released. http://www.tei-c.org/Guidelines/P5/.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.