DATA CURATION NETWORK

Text Encoding Initiative (TEI) Primer

Authors: Courtney Dalton, Cornell University; Emily Kilcer, University at Albany (SUNY); Katie Wampole, Iowa State University; Sarah Swanz, Washington University in St. Louis

Mentor: Mikala Narlock, University of Minnesota, and Sarah Swanz, Washington University in St. Louis

Peer reviewers: Daniel Johnson and John Walsh

Preferred citation: Dalton, Courtney; Kilcer, Emily; Wampole, Katie; and Swanz, Sarah. 2023. Text Encoding Initiative (TEI) Primer. Data Curation Network Github Repository.

Overview

Topic	Description
File Extension	.xml .xsl or .xslt
MIME Type	application/xml *or* text/xml application/xslt+xml or text/xls
Versions	TEI P5
Primary fields or areas of use	Humanities, Social Sciences, Multidisciplinary
Source and affiliation	TEI Consortium
Metadata standards	TEI P5: Guidelines for Electronic Text Encoding and Interchange https://doi.org/10.5281/zenodo.3413524

Preferred citation: Dalton, Courtney; Kilcer, Emily; Wampole, Katie; and Swanz, Sarah. 2023. Text Encoding Initiative (TEI) Primer. Data Curation Network Github Repository.

Key questions for curation review	Is there any TEI customization? Is there any accompanying stylesheet (.xsl or .xslt file)? Is the TEI XML file well-formed and valid against the TEI schema in use? Does the metadata include sufficient documentation of editorial decisions?
Tools for curation review	Text Editor (preferably with syntax highlighting), Oxygen XML Editor (a paid tool with 30-day free trial)
Date Created	September 2023
Created by	Courtney Dalton, cmd347@cornell.edu Emily Kilcer, ekilcer@albany.edu Katie Wampole, kwampole@iastate.edu Sarah Swanz, sswanz@wustl.edu Mentors: Mikala Narlock, mnarlock@umn.edu Sarah Swanz, sswanz@wustl.edu

Table of Contents

Scope

This primer focuses on textual resources or their facsimiles that have been annotated according to Text Encoding Initiative (TEI) conventions. Because TEI is expressed in the Extensible Markup Language (XML), many of the considerations in this primer may be relevant to XML files in general, as well as textual data encoded using other markup languages. In addition, the Music Encoding Initiative (MEI) is based on TEI, and so curation of MEI files can expect to follow a similar process.

Other text corpora, such as machine learning training sets or large language models, are beyond the scope of this primer.

What is the Text Encoding Initiative?

The Text Encoding Initiative (TEI) refers to a set of standards for representing features of a textual document. Key features of TEI documents include the following:

a. Structural Elements: TEI provides a wide range of elements to represent different parts of a text, such as headings, paragraphs, footnotes, lists, tables, stanzas, chapters, etc., as well as document dimensions, markings, and coloring.

b. Metadata: TEI includes descriptive metadata, providing information about the text's title, author, publication date, source, provenance, genre, and so on.

c. Textual Variation: TEI supports encoding variant readings, manuscript transcriptions, and textual corrections, which is useful for studying historical texts with multiple versions.

d. Interpretation and Analysis: TEI enables encoding of features like marginalia, citations, annotations, and interpretations, thereby facilitating in-depth analysis and scholarly research.

e. Linguistic and Semantic Markup: TEI allows for the identification of linguistic features like languages, translations, and semantic elements such as names, places, and dates (named entities).

f. Accessibility and Preservation: TEI-conformant markup helps ensure that texts remain accessible over time and across various digital platforms, aiding in their long-term preservation.

TEI is currently officially expressed in XML. These standards are defined in the *TEI Guidelines*, which are maintained and updated by the TEI Consortium. The *Guidelines* are available on the TEI website or on Zenodo (TEI Consortium. (2022). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Zenodo. https://doi.org/10.5281/zenodo.3413524.

Like XML, TEI is platform independent, extensible, and both human and machine readable. As a markup language, it is designed to formally describe text documents so that they are both human and machine readable, often supporting multiple uses and outputs in subsequent processing.

How and why do researchers use TEI?

TEI is used in digital humanities and digital archives to represent any textual materials, such as poetry, drama,

manuscripts, legislative documents, archival documents, ancient inscriptions, correspondence, and so on. (However, while TEI is focused on textual resources, it is not limited to that.) Scholars might use TEI to create critical scholarly editions, translations, linguistic corpora, historical lexicons, or digital archives.

Using an XSLT transformation, researchers can convert a TEI/XML document into HyperText Markup Language (HTML) and Cascading Style Sheets (CSS) for rendering on the web, to Portable Document Format (PDF) for printing, or to other serialization formats like JavaScript Object Notation (JSON), Resource Description Framework (RDF), and Metadata Object Description Schema (MODS).

Researchers can also use programming languages to extract data from TEI/XML files. For example, a drama scholar might query the TEI file to identify all the characters and determine whether male or female characters have more lines of dialogue.

TEI Examples

Select Collections of TEI documents

- Catalogue of Digital Editions
- Collected Letters of Robert Southey
- EarlyPrint: Curating and Exploring Early Printed English
- Folger Shakespeare Library
- Perseus Digital Library
- Shelley-Godwin Archive

Select Individual TEI files (including sample citations)

Beshero-Bondar, Elisa; Viglianti, Raffaele; Mulligan, Rikk (2020): Frankenstein Variorum - Collations. Carnegie Mellon University. Dataset. https://doi.org/10.1184/R1/11538798.v1

Green, Augustus R. *The Life of the Rev. Dandridge F. Davis, of the African M. E. Church*. Pittsburgh, PA: Ohio A. M. E. Conference, 1850. *Documenting the American South*. 2000. University Library, The University of North Carolina at Chapel Hill. 5 January 2023 <https://docsouth.unc.edu/neh/greena/greena.xml>.

Key questions for curation review

- What source types are being analyzed?
- Are there copyright or licensing restrictions on the text encoded in this dataset ?

- Is the document well-formed and can it be validated against a TEI Schema?
- Has the TEI schema been customized? If so, how?
- Does the metadata include sufficient documentation of editorial decisions?
- Are there other accompanying data types (e.g., images or audio-visual aligned with text) that will be part of the dataset? Do they have any separate copyright or licensing restrictions?
- Is it important to capture the rendered document of the source described in TEI? If so, are there accompanying pdf or HTML/CSS files?

Curation workflow based on the Data Curation Network **CURATED** model

- **Check** the data files and read the documentation

  Verify you can open all of the files in the data set with a text editor. There are a number of different tools for opening and viewing XML files. Any text editor that includes syntax highlighting features (e.g., Notepad++) should be suitable for reviewing content and examining tags.

  Any TEI customization should be accompanied by an .odd file (One Document Does it all). These are also XML files (but commonly given the .odd extension) that can be opened and viewed in a text editor.

  The TEI files may be accompanied by schema files (e.g., .dtd, .rng) or styling sheets (.xslt or .css). If so, make sure you can open those files as well.

- **Understand** the data
  Review the XML files to ensure that they are both *well-formed* and *valid*. XML is *well-formed* when it exhibits correct syntax, such as proper opening and closing tags and nesting of elements. XML is *valid* when it conforms to either a document type definition (DTD) or an XML schema (e.g., Relax NG).

  - Check for character-encoding (UTF-8 or UTF-16) at a minimum and preferably include a language attribute in the header that specifies the language and writing system used. *See* Languages and Character Sets in TEI. TEI files of historical texts may encode non-standard characters like printers' marks.

  - Review XML for well-formedness: You can use Oxygen XML Editor (a paid tool with 30-day free trial) to check well-formedness. A free online alternative to Oxygen XML Editor is the W3C Schools XML Validator to confirm that the XML is well-formed.

  - Validate the file against the schema to identify any errors. The schema must be specified in the header. You can also use Oxygen XML Editor to check validity. There are other free tools for checking well-formedness and validity, such as Jing or xmllint.

In addition to checking well-formedness and validity, standard to any XML file curation, the TEI guidelines specify additional requirements so that the document is considered TEI-conformant:

- Conforms to the TEI Abstract Model: This requires that the TEI Header include a title statement (titleStmt), a publication statement (publicationStmt), and a source statement (sourceDesc).

- Uses the TEI namespace correctly: If the document claims to use the TEI namespace, then the elements used therein must belong to the TEI namespace. (See full TEI element list). Likewise, if the document uses custom elements that do not belong to the TEI namespace, then it should declare an additional namespace at the beginning of the XML document using the xmlns attribute. Searching for the element name in the TEI element list or custom element list should be sufficient verification.

  Using prefixes provides a clear way to differentiate between elements from different namespaces, especially when multiple namespaces are involved in the same document, in order to maintain consistency, readability, and accurate interpretation of the XML content.

  The following is a simplified example of a TEI-conformant XML document that uses the TEI namespace and a custom namespace to declare the elements that belong to each. Note that elements that belong to the TEI namespace use the TEI prefix.

```
<?xml version="1.0" encoding="UTF-8"?>
<tei:TEI xmlns:tei="http://www.tei-c.org/ns/1.0"
xmlns:custom="http://www.example.com/custom-ns">
   <tei:teiHeader>
      <tei:fileDesc>
         <tei:titleStmt>
            <tei:title>Ode to Data Curation Network</tei:title>
            <tei:author>FirstName LastName</tei:author>
         </tei:titleStmt>
         <tei:publicationStmt>
            <tei:publisher>Data Curation Network</tei:publisher>
            <tei:date>2023</tei:date>
         </tei:publicationStmt>
         <tei:sourceDesc>
            <tei:biblStruct>
               <tei:monogr>
                  <tei:title>Book Title</tei:title>
                  <tei:idno type="ISBN">978-1234567890</tei:idno>
               </tei:monogr>
            </tei:biblStruct>
         </tei:sourceDesc>
      </tei:fileDesc>
   </tei:teiHeader>
   <tei:text>
      <tei:body>
```

```
     <tei:div type="poem">
          <tei:lg>
               <tei:l>Roses are red,</tei:l>
               <tei:l>Violets are blue.</tei:l>
               <tei:l>I love DCN,</tei:l>
          <tei:l>And so should you.</tei:l>
          </tei:lg>
     </tei:div>
   <custom:primer>TEI data curation</custom:primer>

     </tei:body>
   </tei:text>
</tei:TEI>
```

- Documented using a TEI-conformant ODD file: Since the TEI Guidelines include hundreds of possible tags, TEI recommends using XML files called ODD (One Document Does it all) files to define the limited set of elements from the TEI that will be used in a particular encoding project. ODD files would also be used to specify any TEI customization.

  Because the TEI Consortium provides a number of basic, general-purpose customizations, such as TEI-Lite, some projects may not include the ODD files in their data submission, but instead refer to these predefined schema by name instead.

  Any TEI project not using one of these general-purpose customizations should include an ODD file specifying its customization. ODD files an be created using the Roma web tool.

For an example of TEI customization files, see the Women Writers Project: https://www.wwp.northeastern.edu/about/methods/customization.html

- **Request** missing information and verify proposed changes
  - Where there are errors in well-formedness, propose modifications for researcher approval (e.g., update tags and nesting).
  - Request character encoding and language attribute, if not supplied.
  - Request schema file or field description for any TEI customization for validation purposes, if not supplied or if not a predefined TEI customization.
  - Identify the invalid schema elements and attributes and request modifications from the researcher for clarity.

- **Augment** the submission
  - A complete TEI file will include a robust TEI Header, which serves as the rough equivalent of a code book found in other disciplines.

    The TEI Header includes the following five components:
    - File description <fileDesc>: This tag should contain the full bibliographical

description of the file itself, similar to a bibliographic citation with author, title, edition, date, etc. It might also contain sponsor and funding information. It will also contain information about the original source(s) from which the electronic text in the TEI file was derived. The child element <sourceDesc> is a container element for bibliographic information such as the origin and publication details of the text.

- Encoding description <encodingDesc>: This tag describes the relationship between an electronic text and its source(s). It might describe how the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, and other editorial decisions.
- Text profile <profileDesc>: This tag documents how the text has been encoded and annotated, including information about the encoding guidelines used, any editorial interventions, and the overall goals of the encoding process. It might also include the language(s) and subject matter, the situation in which it was produced, and the individuals described by or participating in producing it.
- Container element <xenoData>: This tag enables inclusion of non-TEI metadata, such as MARC records for cataloging or Dublin Core.
- Revision history <revisionDesc>: This tag provides a history of changes made during the development of the electronic text, to help with version control or resolve other questions about the history of a file.

Only the first – the file description – is strictly necessary, but we recommend including as many as possible. Information in the TEI Header, especially the <fileDesc>, can often be used to populate a readme.txt file or repository metadata.

- Some researchers include a narrative or description of editorial decisions on a project website. A link to such information can be included in any readme.txt file or even supplied as a text file with the rest of the submission. This information may also be included in the <encodingDesc> tag, specifically the <editorialDecl> tag.

- Update any metadata as needed. Where applicable, link to grant information, source data, etc. Document and include any changes in the curation log. Verify the schema file is included. (Adapted from the DCN Curator checklist.)

- **Transform** (includes preservation actions)
  - TEI files provided as .xml should not need transformation to a more sustainable format.
    Note: If the presentation of the content is significant, consider capturing a PDF or other static output to represent the look and feel of the project.
  - Schema files in ODD, Relax NG, XSLT are also expressed in XML and should not need transformation.
  - Any accompanying files (images, web components, etc.) should be made easily accessible when possible.

- **Evaluate** to make sure this file meets FAIR principles
  - TEI data does not have any specific Evaluate steps. Refer to the DCN CURATE(D) Steps.

- **Document** (what to capture from the curation process)

- TEI data does not have any specific Document steps. Refer to the DCN CURATE(D) Steps.

Bibliography

Text Encoding Initiative Consortium: https://tei-c.org/

TEI Consortium. (2022). TEI P5: Guidelines for Electronic Text Encoding and Interchange (v4.5.0). Zenodo. https://doi.org/10.5281/zenodo.3413524

TEI by Example: https://teibyexample.org/

TAPAS (TEI Archiving Publishing and Access Service): https://www.tapasproject.org/