A Very Short Introduction to TEI EpiDoc

EAJS Winter School
Digital Humanities and Jewish Epigraphy
Utrecht, the Netherlands

Max Grüntgens (instructor, ADWLM)
Thomas Kollatz (instructor, ADWLM, STI)
Ortal-Paz Saar (organizer, UU)

19.–21.2.2018

The paper aims to provide a concise overview to the purpose as well as to the practice of standardized encoding in regard to Humanities sources and data provided by the Text Encoding Initiative (TEI). For the purpose of conciseness as well as clarity – and in concordance with the major topic of the winter school being on history and epigraphy – the paper lays its focus on the TEI subset known as EpiDoc (epigraphic documents). The paper exemplifies the basic structure of a TEI EpiDoc document. It also states best practices in regard to the implementation of the metadata header as well as to the structure of the document body, that incorporates one or more transcription sections and various descriptive as well as interpretative scholarly text sections. There are parts dedicated to the encoding of the process of text constitution, of the most common types of realia (people, places) as well as of various structural and linguistic entities, that may be regarded as semantically more neutral. The paper closes with a troubleshooting section to provide the user with a methodical approach for the solving of difficult textual phenomena, that she may encounter during the encoding.
You may find the *slides here (https://digicademy.github.io/2018_EAJS_WS_1/#/step-1)*.

1 TEI-EpiDoc

TEI-XML is . . .

- . . . ambitious in its **complexity and generality**
- . . . fundamentally no different from that of any other XML markup scheme and . . .
- . . . a **plain text format** and therefore does not rely on a specific or proprietary software suite
- . . . in general conforming to a **simple hierarchic model**
- . . . processible by any general-purpose XML-aware software

Maxim

> ***An encoding makes explicit only those textual features of importance to the encoder.***

TEI-EpiDoc is . . .

- . . . a valid **subset** of the full TEI scheme
- . . . able to handle adequately a **reasonably wide variety** of historical documents (papyri, inscriptions)
- . . . as **small and simple** as is consistent with the other goals

Structure

- TEI root element (`<TEI>`)
- TEI namespace (`<TEI xmlns="http://www.tei-c.org/ns/1.0">`)
- Metadata section (`<teiHeader>`)
- Facsimile section (`<facsimile>`)
- Transcription and general text section (`<text>`)

TEI Header

The TEI header provides **metadata** analogous to that provided by the **title page** of a printed text. It has up to four parts:

1. the **fileDesc**, a **bibliographic description** of the machine-readable text

- full bibliographical description of the computer file itself for proper **citation** or catalogue entry
- information about the **source** from which the file was derived

2. the **encodingDesc** a description of the way it has been **encoded**

- description of **normalization** during transcription
- description of resolved **ambiguities**
- description of level of **analysis**

3. the **profileDesc**, a non-bibliographic description of the text (a **text profile**)

- **classificatory** and contextual information
- it is recommended to enforce a **controlled vocabulary**
- may be of use in any form of **aggregation** or automatic text **processing**

4. the **revisionDesc**, a **revision history**

- provide a history of **changes**
- implements a basic form of **version control**

TEI Text (in EpiDoc)

A TEI text in EpiDoc may be

- unitary (a **single** inscription) or
- composite (a **collection** of single inscriptions on different parts of the monument)
- must have a **body of text**,
  - which in the case of a composite text may consist of different **subdivions** encoded with `<div type="textpart">`,
  - each containing more groups or text sections encoded as **anonymous blocks** with `<ab>` to circumnavigate the specific semantics of paragraphs and alike
- may contain further **elements** as well as global and other **attributes**

2 Central Sections and Elements within the TEI Header

Bibliographic and Legal Information

- `<titleStmt>` contains the `<title>` and a range of other **bibliographic information** about the `<author>` and `<editor>`
- `<publicationStmt>` contains **legal information** concerning the `<authority>` issueing the file and the `<availability>` under a specific `<licence>`

Storage and Materiality

- `<msIdentifier>` **identifies and localizes** the edited object by `<placeName>`, `<repository>`, `<idno>` (identifying number) and so on
- `<physDesc>` keeps metadata related to
  - ... the **materiality** by providing `<objectType>`, `<material>` as well as the spatial `<dimensions>` of the edited object within the `<objectDesc>`
  - ... the individual scribe's or stone cutter's **execution of the text** in the `<handDesc>`
  - ... to the individual **styles of lettering** used within the `<scriptDesc>`
  - ... to the various kinds of **ornamentation** used within the `<decoDesc>`

- `<history>` collects metadata about the **origin and provenance** of the object in question

Corresponding Persons and Languages

- `<listPerson>` within `<particDesc>` collects metadata about the **person and relationships** relating to the document or object
- `<langUsage>` holds **languages** relating to the document or object

3 Central Sections and Elements within the TEI Text

Text Divisions

- `<div>` –> stands for **division** and must take one of six fixed types
  - `type="edition"` brackets the editorial or rather **transcript section**
    - `type="textpart"` brackets distinct **inscription sections** occuring on the same object; this `<div>` must occur within a `type="edition"`
      - `<ab>` is "an **anonymous container** for phrase or inter level elements analogous to, but without the semantic baggage of, a paragraph."; `<ab>` ought to occur only within `type="edition"` or `type="textpart"`
      - `<lb n="0-∞" />` is **self closing** and shows the **beginning of a line**; ought to occur in `<ab>`
      - `<milestone />` is **self closing** and shows **breaks in continuous text**; ought to occur in `<ab>`
  - `type="commentary"` brackets the **commentary section**
    - contains paragraphs `<p>` and further elements
  - `type="apparatus"` brackets the **critical apparatus section**
    - contains further elements associated with **interlinking** of apparatus and transcription
    - **basic apparatus** contains paragraphs `<p>` and further elements
    - **elaborate apparatus** contains a list of apparatus entries `<listApp>` containing `<app>` and `<note>` entries.
    - The apparatus, the lists of apparatus entries and the respective apparatus entries may **interlink** by pointing to their respective sections and lines by usage of `@n` and `@loc`
  - `type="translation"` brackets the **translation section**
    - contains paragraphs `<p>` and further elements
  - `type="bibliography"` brackets the **bibliographic section**

    - ought to contain a **list of bibliographical items**, e.g. `<listBibl>` with `<bibl>` items and further elements

- **Typographic elements** are encoded with `<hi rend="keyword">`

Realia

- **Persons** or a person's name ought to be encoded with `<persName>`; may include further subdivisions like `<forename>`, `<surename>`
- **Ethnic groups** may be encoded with `<orgName>`
- **Places** or place names (settlements) may be encoded with `<placeName>`
- **Geographic features** like mountains and rivers may be encoded with `<geogName>`
- The realia may be **interlinked** and **standardized** via use of
  - `@type` do distinguish several **specific taxonomic types** of f.e. place names, e.g. settlements, fortifications, hamlets and solitary farms.
  - `@nymRef` which will contain a URL or other URI pointing to the **standard form** of this name (nominative singular; normalized spelling)
  - `@ref` to point to a place identifier at a local database or online **gazetteer**
- an **alternative encoding** of a general purpose name or referring string or phrase may be supplied by `<rs>`
  - `@type` and `@subtype` attributes clarifying the **specific function** based on a controlled vocabulary may be supplied by the encoder
  - the encoder may further point to a **standardized form** in the header or an external file

Text Constitution (Leiden Brackets)

See also **Leiden Cheatsheet** with an overview of Panciera and Leiden.

Divisions

- **Line** breaks `<lb n="1"/>begin of first line <lb n="2"/>begin of second line`
- **Word divided** across lines `<lb n="1"/> abc <lb n="2" break="no"/> def`
- Text **divisions** `<div type="textpart" subtype="face" n="r"> ... </div> <div type="textpart" subtype="face" n="v"> ... </div>`

Readability

- Clearly **readable** text `abc`

- Clear but **incomprehensible** letters `<orig> abc </orig>`
- Letters **ambiguous** outside of their context `<unclear> abc </unclear>`
- Vestiges of letters visible but **illegible** `<gap reason="illegible" quantity="3" unit="character"/>`
- Text visible to previous editor, but **now lost** `<supplied reason="undefined" evidence="previouseditor"> abc </supplied>`
- Text **restored** by comparison with parallel copy `<supplied reason="undefined" evidence="parallel"> abc </supplied>`

Text Execution

- **Apices** `<hi rend="apex"> a </hi>`
- **Supralinear** lines `<hi rend="supraline"> abc </hi>`
- **Ligatured** letters `<hi rend="ligature"> ab </hi>`

Erasures

- **Erased** `<del rend="erasure"> abc </del>`
- Erased and **lost** `<del rend="erasure"><gap reason="lost" quantity="3" unit="character"/></del>`
- Text **struck over** erasure `<add place="overstrike"> abc </add>`
- Erased and overstruck or **corrected** `<subst><del rend="corrected"> ab </del><add place="overstrike"> cd </add></subst>`

Additions

- Text **added above** by ancient hand `<add place="above"> abc </add>`
- Text **added below** by ancient hand `<add place="below"> abc </add>`
- Characters lost but **restored** `<supplied reason="lost"> abc </supplied>`
- Characters **restored tentatively** `<supplied reason="lost" cert="low"> abc </supplied>`
- Word **incompletely restored** `<w part="I"> abc <supplied reason="lost"> de </supplied></w>`
- **Subaudible word** suppled by editor `<supplied reason="subaudible"> abc </supplied>`

Loss and Deficiencies

- **Characters lost**, lacuna `<gap reason="lost" quantity="3" unit="character"/>`
- Lacuna, **extent unknown** `<gap reason="lost" extent="unknown" unit="character"/>`

- Lacuna, **approximate extent** `<gap reason="lost" quantity="5" unit="character" precision="low"/>`
- Lacuna, **praenomen** `<name> <gap reason="lost" atLeast="1" atMost="3" unit="character"/> </name>`
- Lacuna, range of **possible extent** `<gap reason="lost" atLeast="5" atMost="7" unit="character"/>`

Lost Lines

- Line **lost** `<gap reason="lost" quantity="1" unit="line"/>`
- Line(s) lost at start or end of text `<gap reason="lost" extent="unknown" unit="line"/>`
- Line **possibly lost** `<gap reason="lost" quantity="1" unit="line"> <certainty match=".." locus="name"/> </gap>`
- Line(s) possibly lost at start or end of text `<gap reason="lost" extent="unknown" unit="line"> <certainty match=".." locus="name"/> </gap>`

Modern Corrections

- Superfluous letters; **suppressed** by editor `<surplus> abc </surplus>`
- Omitted letters; **added** by editor `<supplied reason="omitted"> abc </supplied>`
- Words **omitted** by editor for brevity `<gap reason="ellipsis"/>`
- Letters **corrected** by editor `<choice> <corr> ab </corr> <sic> bc </sic> </choice>`
- Word **regularized** by editor `<choice> <reg> ab </reg> <orig> cd </orig> </choice>`
- Editor's **note**, i.e. 'sic' `<note>!</note>, <note>sic</note>, <note>e.g.</note>`

Abbreviations

- **Expansion** of abbreviation `<expan> <abbr> a </abbr> <ex> bc </ex> </expan>`
- **Tentative expansion** of abbreviation `<expan> <abbr> a </abbr> <ex cert="low"> bc </ex> </expan>`
- **Incomplete expansion** of abbreviation `<w part="I"> <expan> a <ex> bc </ex> </expan> </w>`
- Incomplete expansion (**Duke**) `<expan> a <ex> bc </ex> </expan>`
- Abbreviation: **expansion unknown** `<abbr> a </abbr>`
- Expansion of **symbol** `<expan> <ex> ab </ex> </expan>`
- `<am>` indicates an abbreviation **marker** (punctuation, characters)

Spaces and Gaps

- Text **not completed** by stonecutter `<gap reason="omitted" extent="unknown" unit="character"/>` or `<gap reason="omitted" extent="unknown" unit="line"/>`
- **Space** left on stone `<space quantity="1" unit="character"/> <space quantity="3" unit="character"/>`
- Space on stone, **extent unknown** `<space extent="unknown" unit="character"/>`

Text Direction

- **Direction** of text `<lb n="1" rend="right-to-left"/>`; for boustrophedon

Numerals

- Numeral (**Roman**) `<num value="12"> XII </num>`
- Numeral (**Greek**) `<num value="1">` α `</num>`
- Numeral (**Hebrew**) `<num value="1" type="hebrew">` א `</num>`

Symbols

- **Symbols** `<g type="leaf"/>`; usage of **controlled vocabulary** is recommended

Global attributes

- `@xml:id` → unique **identifier** e.g. hierarchic structure of the document, cryptographic hashes, ...
  - The attribute's value must be necessarily **unique** within the document scope. Uniqueness within corpus scope is strongly adviced.
  - The attribute may be used to provide **anchor points** wherever necessary to which f.e. apparatus entries may point.
- `@n` → **mnemonic label** e.g. Stephanus' pagination, ...
  - The attribute's value needs not necessarily be unique within the scope of the document or the corpus.
  - The attribute may be used to provide **anchor points** wherever necessary to which f.e. apparatus entries may point.
- `@xml:lang` → **language** tag e.g. de, en, fr, ...
- `@rend` → **typography**, orientation, e.g. bold, italic, ...

Other Useful Elements

- `<term>` contains a word or phrase regarded as a **technical term**. It can carry `@type` and `@subtype` attributes and may `@ref` to a glossary or index.
- **Dictionary** or **glossary** style text parts may be modeled by alternating `<term>` and `<gloss>` or semantically neutral `<seq>`.
- `<foreign>` contains a word or a phrase in a **foreign language** other than the surrounding text. It ought to carry a `@xml:lang` attribute.
- `<handShift>` indicates a shift of the **writing hand**. It ought to point to a `<handNote>` in the `<teiHeader>` via a `@new` attribute.
- `<quote>` and `<bibl>` may be combined – as needed – with a surrounding `<cit>` to refer to **citations** and **quotations**.
- `<seg>` may be used to represent any **text segment** for whatever purpose deemed necessary.
- `<w>` may be used to represent any **word segment** (often for linguistic purposes). It may take a `@lemma` attribute.
- `<c>` may be used to represent any **character** (often for linguistic purposes). It may take various attributes.
- `<pc>` may be used to represent any **punctuation mark**. The exact purpose of the mark may be given with `@type` and `@subtype` attributes.

4 Troubleshooting

What to do, when I am stuck while encoding an inscription?

1. Explicate the textual feature you want to describe and its relevance to your encoding as well as to your research question in simple language:
   - Are there similar or related textual phenomena you have already encoded?
   - Is the textual phenomenon specific or universal?
   - What's the encoding's purpose in regard to your research question?
2. Reread the relevant guideline chapters as well as the respective reference tags' reference pages of the related and already encoded text features.
   - Do you need a specific, semantically charged tag to encode the textual phenomenon?
   - May the textual phenomenon be encoded with a semantically neutral tag specified by attributes?
   - Search the mailing list by utilizing the answers made in regard to the questions stated above.
3. State your problem on the mailing list by providing ...
   - ... a minimal example of the textual phenomenon in its text surroundings.
   - ... the purpose of your desired encoding.
   - ... a concise statement elaborating why the available encoding proposals do not suit your needs.
   - ... (optional) an encoding proposal based on your minimal example.