Introduction to TEI

Magdalena Turska

Around TEI in 90 minutes

Basic structure and core TEI elements

- TEI Universe
- XML foundations
- TEI Core

What is text? (and what are we doing when we are editing?)

What is text? I am not so nave as to imagine that question could ever be finally settled. Asking such a question is like asking How long is the coast of England?.

J. McGann

Text is what you look at. And how you look at it.

P. Sahle

Puzzle

RhymeWhatinspiredthisamorousrhyme?Twopartsvodkaonepartlime

Puzzle 2

Rhyme What inspired this amorous rhyme? Two parts vodka one part lime

Puzzle 3

Rhyme What inspired this amorous rhyme? Two parts vodka one part lime

Various texts

- Poem: If
- Book: Treasure Island
- Play: Romeo & Juliet
- Letter: Robert Graves to William Graves

The ambivalence continues

- every project is different and we delight in flexibility and customization
- yet we aim at standardization & interoperability

(...) documents worth encoding (...) are very different from customer letters. But not that different, and eight out of ten probably will benefit from staying within the confines of a well thought-out standard schema and its surrounding processing rules. And even the two that dont may benefit from staying within that standard schema as far as possible. - M. Mueller

Common denominators

- basic metadata: title, authors and other contributors, publication info
- logical text structure: divisions, paragraphs, strophes, speeches or verses
- physical text structure: pages, columns and lines
- highlights: hi, emph, foreign, title etc
- regularization, abbreviations and corrections
- people and places
- dates

... but the list never ends and depth of encoding goes as far as the coast of England is long

The long haul

Presentation by its very nature is ephemeral: bound to be modified over time or multiplied for other use scenarios

Computer stuff gets obsolete pretty quickly: a decade or two or maybe 2 years if you're unlucky

Need to refurbish the interfaces, migrate the underlying software and perhaps data as well is one constant in long-term preservation

Its the data source, the encoding that is of long-lasting value and not the presentation

Good, formal, well documented data model will lend itself to fairly easy conversion into any new solution that comes

What are our hopes and promises?

(not only when applying for funding)

- scholarly use
- and RE-use

- interchange
- long-term preservation

What ensures the best chance of long-term preservation and reuse is the concentration on openly licensed, quality encoding conforming as best as possible to standards.

The markup language we use must be able to ...

- specify all the characters found
- make explicit the structures perceived
- represent that structure in a linear processable form
- additionally supply a variety of metadata or annotations

XML is a good fit (for the most part...)

Community standard: TEI

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.

TEI Guidelines: encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.

Widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation.

The TEI Consortium is a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors. In commemoration of the TEI community's 30th anniversary, it will be awarded the 2017 Antonio Zampolli Prize from the Alliance of Digital Humanities Organizations.

`tei-community.png`

Join the TEI, they said...

TEI is often called a de facto standard, likely because it's a community of users, developing from the bottom up.

If you have questions about anything broadly connected with TEI ask at TEI-L@listserv.brown.edu; you get answers not only from us, but TEI experts around the world. Don't be shy: questions from those of all levels of ability stop the list becoming too technical, everyone benefits from having the answers be public – and you benefit by reading (and sometimes answering!) others' problems soldier.jpg

TEI encoding scheme is expressed in XML

The rules and recommendations made in TEI Guidelines are expressed in terms of what is currently the most widely-used markup language for digital resources of all kinds: the Extensible Markup Language (XML), as defined by the World Wide Web Consortium's XML Recommendation. However, the TEI encoding scheme itself does not depend on this language; it was originally formulated in terms of SGML (the ISO Standard Generalized Markup Language), a predecessor of XML, and may in future years be re-expressed in other ways as the field of markup develops.

TEI document structure

Structure of a TEI Document

There are two basic structures of a TEI Document:

- TEI (TEI document) contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a teiCorpus element.
- teiCorpus contains the whole of a TEI encoded corpus, comprising a single corpus header and one or more TEI elements, each containing a single text header and a text.

TEI basic structures (1)

¡TEI¿ ¡teiHeader¿¡!– required –¿¡/teiHeader¿ ¡facsimile¿¡!– optional–¿¡/facsimile¿ ¡sourceDoc¿¡!– optional –¿¡/sourceDoc¿ ¡text¿¡!– required if no facsimile or sourceDoc– ¿¡/text¿ ¡/TEI¿

TEI basic structures (2)

¡teiCorpus¿ ¡teiHeader¿¡!– required –¿¡/teiHeader¿ ¡TEI¿¡!– required –¿¡/TEI¿ ¡!– More ¡TEI¿ elements –¿ ¡/teiCorpus¿

A corpus is a collection of text and header pairs that also has its own header.

What is a text?

- A text may be unitary or composite
  - unitary: forming an organic whole
  - composite: consisting of several components which are in some important sense independent of each other

Unitary text structure

a unitary text contains

- optional front matter
- body (required)
- optional back matter

¡text¿ ¡front¿¡!– optional –¿¡/front¿ ¡body¿¡!– required –¿¡/body¿ ¡back¿¡!– optional –¿¡/back¿ ¡/text¿

Composite text structure

A composite text contains

- optional front matter
- group with text inside (required)
- optional back matter

¡text¿ ¡front¿¡!– front matter to an anthology –¿¡/front¿ ¡group¿ ¡text¿ ¡body¿ ¡p¿ ¡!– content of first text –¿¡/p¿ ¡/body¿ ¡/text¿ ¡!– more texts here –¿ ¡/group¿ ¡back¿¡!– back matter to an anthology –¿¡/back¿ ¡/text¿

Another Grouped Text Example

¡group¿ ¡text¿ ¡!– optional front matter –¿ ¡body¿¡!– First Body –¿ ¡/body¿ ¡!– optional back matter –¿ ¡/text¿ ¡text¿ ¡!– optional front matter –¿ ¡body¿ ¡/body¿ ¡!– optional back matter –¿ ¡/text¿ ¡/group¿

’Core’ elements

The so-called ’Core’ module groups together elements which may appear in any kind of text and the tags used to mark them in all TEI documents. This includes:

- divisions
- paragraphs
- simple verse and drama structures
- highlighting, emphasis and quotation
- simple editorial changes
- basic names numbers, dates, addresses
- simple links and cross-references
- lists, notes, annotation, indexing
- graphics
- reference systems, bibliographic citations

Divisions

div (division) marks logical document hierarchy

- Fundamental unit for any kind of text
- div can appear directly inside body or nested in other div (divisions)
- @type attribute is usually used to distinguish division types, e.g. volume, book, chapter, section
- headchild is used to provide division heading

¡div¿ ¡head¿Vorrede¡/head¿

¡p¿ Die menchliche Vernunft hat das beondere Schickal in einer Gattung ihrer Erkentnie: da ie durch Fragen belatigt wird, die ie nicht a weien kan; denn ie ind ihr durch die Natur der Vernunft elbt aufgegeben, die ie aber auch nicht beantworten kan, denn ie uberteigen alles Vermogen der menchlichen Vernunft. ¡/p¿ ¡/div¿

Paragraphs

p (paragraph) marks paragraphs in prose

- Fundamental unit for prose texts
- p can contain all the phrase-level elements in the core
- p can appear directly inside body or inside div (divisions)

¡p¿ Thanks for yours of this morning. I hope ¡lb/¿you have had my card posted last Monday. ¡lb/¿On Mond. next I lecture the ¡orgName ref=”Fieldclub”¿Field Club¡/orgName¿ - ¡lb/¿a Nat. Hist. Association, in the lines of our ¡lb/¿old Society - Geological, (you + me) + Botanical ¡lb/¿(New) Do you remember: you¡supplied¿r¡/supplied¿ old ¡lb/¿Black Molt? ¡/p¿

Simple Verse

¡lg type=”stanza”¿ ¡l¿It seemed that out of battle I escaped¡/l¿ ¡l¿Down some profound dull tunnel, long since scooped¡/l¿ ¡l¿Through granites which titanic wars had groined.¡/l¿ ¡/lg¿

Simple Drama

¡sp¿ ¡speaker¿The reverend Doctor Opimiam¡/speaker¿ ¡p¿I do not think I have named a single unpresentable fish.¡/p¿ ¡/sp¿ ¡sp¿ ¡speaker¿Mr Gryll¡/speaker¿ ¡p¿Bream, Doctor: there is not much to be said for bream.¡/p¿ ¡/sp¿

Highlighting

By highlighting we mean the use of any combination of typographic features (font, size, hue, etc.) in a printed or written text in order to distinguish some passage of a text from its surroundings. For words and phrases which are:

- distinct in some way (e.g. foreign, archaic, technical)
- emphatic or stressed when spoken
- not really part of the text (e.g. cross references, titles, headings)
- a distinct narrative stream (e.g. an internal monologue, commentary)
- attributed to some other agency inside or outside the text (e.g. direct speech, quotation)
- set apart in another way (e.g. proverbial phrases, words mentioned but not used)

Highlighting Examples

- hi (general purpose highlighting);distinct (linguistically distinct) ¡p¿Last week I wrote (to order) a strong ¡lb/¿bit of Blank: on ¡hi rend=”ul”¿Antaeus v. Heracles¡/hi¿. ¡lb/¿These are the best lines, methinks: ¡lb/¿(N.B. Antaeus deriving strength from his Mother Earth ¡lb/¿nearly licked old ¡distinct¿Herk¡/distinct¿.) ¡/p¿
- Other similar elements include: emph, mentioned, soCalled, term and gloss

Quotation

Quotation marks can be used to set off text for many reasons, so the TEI has the following elements:

- q (separated from the surrounding text with quotation marks)
- said (speech or thought)
- quote (passage attributed to an external source)
- cit (groups a quotation and citation)

¡cit¿ ¡quote¿ ¡l¿... How Earth herself empowered him with her trick,¡/l¿ ¡l¿Gave him the grip and stringency of Winter,¡/l¿ ¡l¿And all the ardour of th’ invincible Spring;¡/l¿ ¡/quote¿ ¡bibl¿ ¡author¿Wilfred Owen¡/author¿ ¡title ref=”works.xmlWO123”¿Letter to Leslie Gunston / The Wrestler¡/title¿ ¡date when=”1917-07”¿July 1917¡/date¿ ¡/bibl¿ ¡/cit¿

Simple Editorial Changes: choice and Friends

- choice (groups alternative editorial encodings)
- Errors:
  - sic (apparent error)
  - corr (corrected error)
- Regularization:
  - orig (original form)
  - reg (regularized form)
- Abbreviation:
  - abbr (abbreviated form)
  - expan (expanded form)

Choice Example

¡p¿...any might, ¡unclear reason=”scribbled”¿majesty¡/unclear¿, ¡choice¿ ¡abbr¿domin¡/abbr¿ ¡expan¿domin¡ex¿ion¡/ex¿¡/expan¿ ¡/choice¿ or power...¡/p¿

Additions, Deletions, and Omissions

- add (addition to the text, e.g. marginal gloss)
- del (phrase marked as deleted in the text)
- gap (indicates point where material is omitted)
- unclear (contains text unable to be transcribed clearly)

Example of add, del, gap, and unclear

¡p¿ ¡add place=”left”¿My ¡/add¿ ¡del rend=”stroked”¿It’s ¡/del¿ ¡add place=”above”¿ ¡del rend=”stroked”¿The ¡/del¿ ¡/add¿ subject ¡del rend=”stroked”¿of¡/del¿ is War, and the ¡unclear¿pity ¡/unclear¿of ¡del rend=”stroked”¿it¡/del¿ War. ¡lb/¿ The Poetry is in the pity. ¡/p¿

Basic Names

- name (a name in the text, contains a proper noun or noun phrase)
- rs (a general-purpose name or referencing string)

The @type attribute is useful for categorizing these, and they both also have @key, @ref, and @nymRef attributes.

Basic Numbers and Measures

- num (marks a number of any sort)
- measure (marks a quantity or commodity)
- measureGrp (groups specifications relating to a single object)
- While num has simple @type and @value attributes, measure has @type, @quantity, @unit and @commodity attributes

Number and Measure examples

¡l¿With a ¡num value=”1000”¿thousand¡/num¿ pains that vision’s face was grained;¡/l¿ ... only ¡measure type=”distance” unit=”m” quantity=”3218.69”¿two miles¡/measure¿ from the front....

Dates

- date (contains a date in any format and includes a @when attribute for a regularised form and a @calendar attribute to specify what calendar system)
- time (contains a time in any format and includes a @when attribute for a regularised form)

¡date when=”1917-07”¿July 1917.¡lb/¿ Wednesday¡/date¿

Simple Linking

- ptr (defines a pointer to another location)
- ref (defines a reference to another location, with optional linking text)
- Both elements have:
  - @target attribute taking a URI reference
  - @cRef attribute for canonical referencing schemes
- If the linking text is able to be generated, ptr and ref might be used in the same place.

Simple Linking Example

See ¡ref target=”Section12”¿section 12 on page 34¡/ref¿.

See ¡ptr target=”Section12”/¿.

Lists

- list (a sequence of items forming a list)
- item (one component of a list)
- label (label associated with an item)
- headLabel (heading for column of labels)
- headItem (heading for column of items)

Simple List Example

The previous slide contained only:

¡div¿ ¡head¿Lists¡/head¿ ¡list¿ ¡item¿list (a sequence of items forming a list)¡/item¿ ¡item¿item (one component of a list)¡/item¿ ¡item¿label (label associated with an item)¡/item¿ ¡item¿headLabel (heading for column of labels)¡/item¿ ¡item¿headItem (heading for column of items)¡/item¿ ¡/list¿

¡/div¿

Notes

- note (contains a note or annotation)
- Notes can be those existing in the text, or provided by the editor of the electronic text
- A @place attribute can be used to indicate the physical location of the note
- Notes should usually be encoded where its identifier/mark first appears; notes can also be kept separately and point back to their location with a @target attribute

¡note¿Painted by ¡persName¿John Singer Sargent¡/persName¿, 1918¡/note¿

Graphics

- graphic (indicates the location of an inline graphic, illustration, or figure)
- binaryObject (encoded binary data embedding a graphic or other object)
- The figure module provides figure and figDesc for more complex graphics with figDesc

¡figure¿ ¡graphic url=”materials/postcard-front.jpg”/¿ ¡figDesc¿A postcard image of two men relaxing at a table, smoking pipes and drinking. A dog and potted fruit tree are nearby with a house over the wall in the distance.¡/figDesc¿ ¡/figure¿

XML

Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable.

Defined by the W3C's XML 1.0 Specification and by several other related specifications, all of which are free open standards.

Strictly speaking, XML is a metalanguage, that is, a language used to describe other languages (or, as we often say: vocabularies). TEI is one of the existing vocabularies expressed in XML.

UNICODE

XML document is a string of characters.

Almost every legal Unicode character may appear in an XML document.

Markup and content

The characters making up an XML document are divided into markup and content, which may be distinguished by the application of simple syntactic rules.

- markup either begins with ¡ and ends with a ¿
- or begins with the character & and ends with a ;
- strings of characters that are not markup are content

Tag

A markup construct that begins with ¡ and ends with ¿.

Tags come in three flavors:

- start-tags; for example: ¡div¿
- end-tags; for example: ¡/div¿
- empty-element tags; for example: ¡lb/¿

Element

A logical document component which either begins with a start-tag and ends with a matching end-tag or consists only of an empty-element tag.

The characters between the start- and end-tags, if any, are the element's content, and may contain markup, including other elements, which are called child elements. An example of an element is ¡salute¿Hello, world.¡/salute¿

. Another is ¡lb/¿

.

Attribute

A markup construct consisting of a name/value pair that exists within a start-tag or empty-element tag. In the example the element img has two attributes, src and alt: ¡img src=”madonna.jpg” alt=’Foligno Madonna, by Raphael’/¿

Another example would be ¡step number=”3”¿Connect A to B.¡/step¿

where the name of the attribute is number and the value is 3.

Attributes

An XML attribute can only have a single value and each attribute can appear at most once on each element.

Where a list of multiple values is desired, this must be done by encoding the list into a well-formed XML attribute with some format beyond what XML defines itself. Usually this is either a comma or semi-colon delimited list or, if the individual values are known not to contain spaces, a space-delimited list can be used.

¡div class=”inner greeting-box”¿Hello!¡/div¿

where the attribute class has both the value inner greeting-box and also indicates the two CSS class names inner and greeting-box.

XML declaration

XML documents may begin by declaring some information about themselves, as in the following example:

¡?xml version=”1.0” encoding=”UTF-8”?¿

Well-formedness

XML document must be a well-formed text it needs to satisfy a list of syntax rules provided in the specification.

- The document contains only properly encoded legal Unicode characters
- None of the special syntax characters such as ¡ and & appear except when performing their markup-delineation roles
- The begin, end, and empty-element tags that delimit the elements are correctly nested, with none missing and none overlapping
- The element tags are case-sensitive; the beginning and end tags must match exactly.
- A single root element contains all the other elements.
- Tag names cannot contain any of the characters !”#$%&’()*+,/;¡=¿?@[\]ˆ‘{—}˜, nor a space character, and cannot start with -, ., or a numeric digit.

DOM

The W3C Document Object Model (DOM) is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure, and style of a document. The XML DOM defines the objects and properties of all XML elements, and the methods (interface) to access them.

The DOM says:

- The entire document is a document node
- Every XML element is an element node
- The text in the XML elements are text nodes
- Every attribute is an attribute node
- Comments are comment nodes
- Text is Always Stored in Text Nodes

A common error is to expect an element node to contain text. However, the text of an element node is stored in a text node. In this example: ¡year¿2005¡/year¿, the element node ¡year¿¡/year¿, holds a text node with the value ”2005”. ”2005” is not the value of the ¡year¿¡/year¿ element!

Node Parents, Children, and Siblings

The nodes in the node tree have a hierarchical relationship to each other. The terms parent, child, and sibling are used to describe the relationships.

- Parent nodes have children.
- Children on the same level are called siblings
- In a node tree, the top node is called the root
- Every node, except the root, has exactly one parent node
- A node can have any number of children
- A leaf is a node with no children
- Siblings are nodes with the same parent

Node Types

Different W3C World Wide Web Consortium node types and descriptions:

- Document represents the entire document (the root-node of the DOM tree)
- DocumentFragment represents a "lightweight" Document object, which can hold a portion of a document
- DocumentType provides an interface to the entities defined for the document
- ProcessingInstruction represents a processing instruction
- EntityReference represents an entity reference
- Element represents an element
- Attr represents an attribute
- Text represents textual content in an element or attribute
- CDATASection represents a CDATA section in a document (text that will NOT be parsed by a parser)
- Comment represents a comment
- Entity represents an entity
- Notation represents a notation declared in the DTD

Namespaces

- Namespace name is an identifier given to an XML vocabulary
- It looks a lot like URI, eg. TEI namespace name is http://www.tei-c.org/ns/1.0
- It doesn't mean that this URI actually points to something (like a schema), this is just a name!

XML namespaces are used for providing uniquely named elements and attributes in an XML document. An XML instance may contain element or attribute names from more than one XML vocabulary. If each vocabulary is given a namespace, the ambiguity between identically named elements or attributes can be resolved.

Mental shortcut to namespaces

If you assume that each XML vocabulary is basically a language, then namespace is a name of this language.

So saying that a particular element comes from a given namespace resolves the ambiguity, just like with natural languages, eg. polish:fart means "luck" and is not smelly at all, contrary to english:fart spanish:hola meaning "hi!" is not the same as polish:hola meaning "hey, man, now wait a minute!"

Back to XML html:div is not quite the same as tei:div, but they could be used in the same document

How to declare a namespace for a document?

use xmlns attribute on root element, eg:

¡TEI xmlns=”http://www.tei-c.org/ns/1.0”¿

All descendants inherit the namespace from their parent, so no need to repeat it on all other elements

How to mix elements from different namespaces?

use xmlns attribute on elements from ’embedded’ namespaces, eg: ¡include xmlns=”http://www.w3.org/20
href=”elementsummary.xml”/¿

declare namespace up top, give it a prefix and use prefixed element names, eg: ¡TEI xmlns=”http://www.tei-c.org/ns/1.0” xmlns:skos=”http://www.w3.org/2004/02/skos/core#” xml:lang=”en”¿ .... somewhere down in the document .... ¡skos:symbol¿blah¡/skos:symbol¿

If you don’t put xmlns attribute anywhere, all elements belong to the default, empty namespace!

¡?xml version=”1.0” encoding=”UTF-8”?¿ ¡tei:bibl xmlns:tei=”http://www.tei-c.org/ns/1.0”¿ ¡title¿eXist¡/title¿ ¡x:author xmlns:x=”http://www.tei-c.org/ns/1.0”¿Adam Retter¡/x:author¿ ¡author¿Erik Siegel¡/author¿ ¡publisher¿O’Reilly¡/publisher¿ ¡date¿2014¡/date¿ ¡idno type=”ISBN”¿9781449337100¡/idno¿ ¡/tei:bibl¿

All elements above are in the TEI namespace

Different prefixes do not matter! tei:author and x:author are the same