

Introduction to TEI XML

• • •

James Cummings and Christopher Ohge

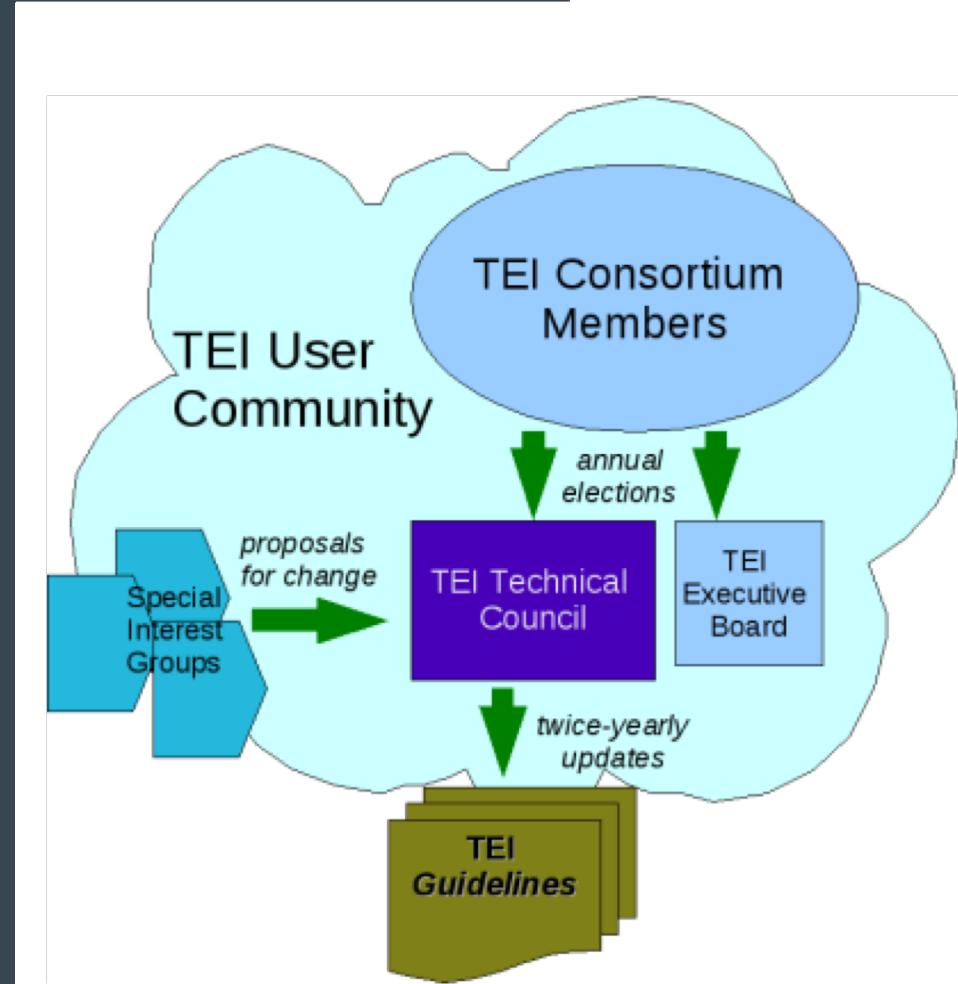
@jamescummings, James.Cummings@newcastle.ac.uk

@cmohge, christopher.ohge@sas.ac.uk

What is the TEI?

- An international consortium of institutions, projects and individual members (established in 1987)
- A community of users and volunteers creating hardware- and software-independent methods for encoding and archiving humanities data
- A flexible set of guidelines with recommendations and examples of over 570 markup distinctions, usually expressed in XML
- A tool for producing customized schemas for validating your project's digital texts
- A set of free and openly-licensed stylesheets for transformations to many formats (e.g. HTML, Word, PDF, Databases, RDF/LinkedData, Slides, ePub, etc.)
- A consensus-based way of organizing and structuring textual resources, images, and other media
- An archival format for long-term preservation of digital data

- # Parts of the TEI Community
- TEI Consortium Members
 - TEI Executive Board
 - TEI Technical Council
(which produces the Guidelines and Software)
 - TEI User Community
 - TEI Special Interest Groups



What is XML?

- XML is a **metalinguage** (a language for how to describe another language), which allows you the flexibility to design your own XML rules.
- It is a **formal model** that represents texts in an **ordered hierarchy**. We think and write with hierarchical structures, and XML is a practical attempt to formalise and represent documents in a machine-readable language.
- An **ordered hierarchical content object** (OHCO): The hierarchy imposed on documents depends on the state of surviving documents as well as one's research questions. The same document can be encoded in more than one hierarchical (or non-hierarchical, for that matter) structure.
- Computers operate more quickly and efficiently on **trees** (ordered hierarchies) than they can on nonhierarchical (flat) documents. Large amounts of data can be managed and transformed efficiently if a document is modelled as a tree.

What is the difference between XML and TEI?

- The XML specification details the rules for the language (e.g. that elements are delineated with angle brackets)
- There are hundreds of other vocabularies of XML in every domain of knowledge you might imagine.
- Journal publishing uses JATS XML, ebooks use BITS XML, e.g.
- There is even a BeerXML standard.
- The TEI Guidelines recommend a specific vocabulary (e.g. the names of elements and attributes) and structure.

Example: A play encoded in TEI XML

SCENE I. On a ship at sea ; a tempestuous noise of thunder and lightning heard.

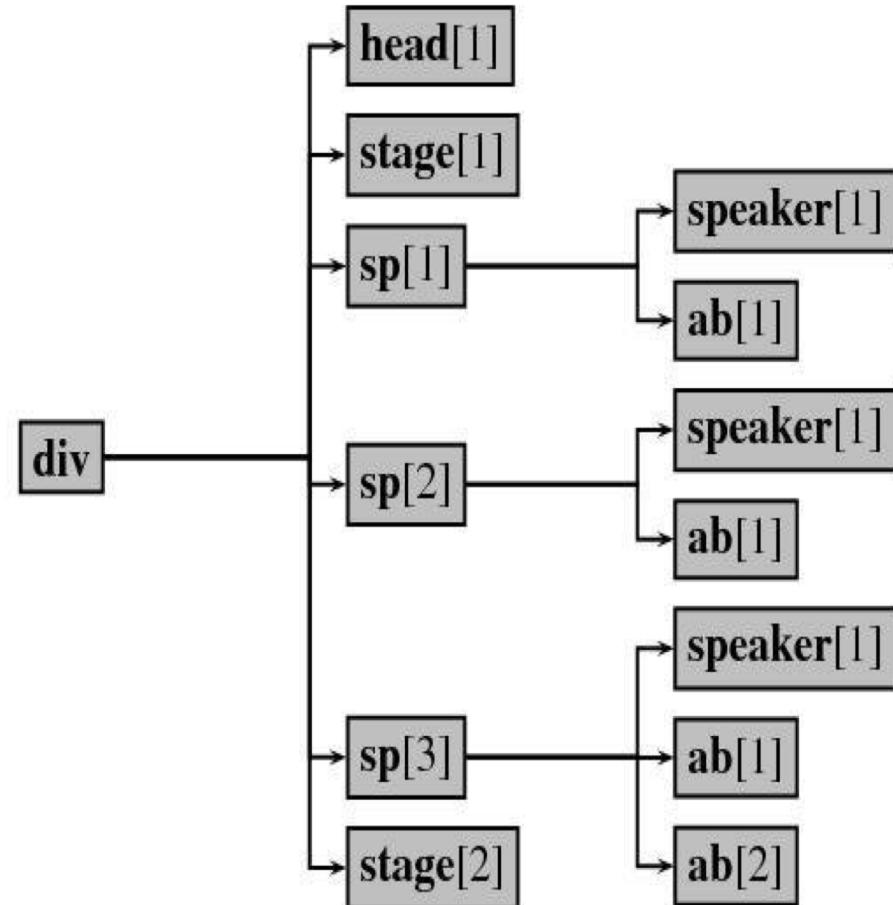
Enter a Shipmaster and a Boatswain.

Master. Boatswain !
Boats. Here, master ; what cheer ?
Master. Good ! Speak to th' mariners ; fall to 't yarely, or we run ourselves aground ; bestir, bestir. [Exit.]

```
<div n="1.1">
  <head>
    SCENE I. On a ship at sea: a tempestuous
    noise of thunder and lightning heard.
  </head>
  <stage>Enter a Master and a Boatswain</stage>
  <sp who="#master">
    <speaker>Master</speaker>
    <ab>Boatswain!</ab>
  </sp>
  <sp who="#boatswain">
    <speaker>Boatswain</speaker>
    <ab>Here, master: what cheer?</ab>
  </sp>
  <sp who="#master">
    <speaker>Master</speaker>
    <ab>Good, speak to the mariners: fall to't,
      yarely,</ab>
    <ab>or we run ourselves aground: bestir,
      bestir.</ab>
  </sp>
  <stage>Exit</stage>
</div>
```

```
<div n="1.1">
  <head>
    SCENE I. On a ship at sea: a tempestuous
    noise of thunder and lightning heard.
  </head>
  <stage>Enter a Master and a Boatswain</stage>
  <sp who="#master">
    <speaker>Master</speaker>
    <ab>Boatswain!</ab>
  </sp>
  <sp who="#boatswain">
    <speaker>Boatswain</speaker>
    <ab>Here, master: what cheer?</ab>
  </sp>
  <sp who="#master">
    <speaker>Master</speaker>
    <ab>Good, speak to the mariners: fall to't,
      yarely,</ab>
    <ab>or we run ourselves aground: bestir,
      bestir.</ab>
  </sp>
  <stage>Exit</stage>
</div>
```

Simplified rendering as a tree



<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>



P5 Guidelines — English ▾ Search

P5: Guidelines for Electronic Text Encoding and Interchange

Version 3.0.0. Last updated on 29th March 2016, revision 89ba24e

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]



Front Matter

Title

- I. Releases of the TEI Guidelines
- II. Dedication
- III. Preface and Acknowledgments
- IV. About These Guidelines
- V. A Gentle Introduction to XML
- VI. Languages and Character Sets

Back Matter

- Appendix A Model Classes
- Appendix B Attribute Classes
- Appendix C Elements
- Appendix D Attributes
- Appendix E Datatypes and Other Macros
- Appendix F Bibliography
- Appendix G Prefatory Notes
- Appendix H Colophon

Text Body

- 1 The TEI Infrastructure
- 2 The TEI Header
- 3 Elements Available in All TEI Documents
- 4 Default Text Structure
- 5 Characters, Glyphs, and Writing Modes
- 6 Verse
- 7 Performance Texts
- 8 Transcriptions of Speech
- 9 Dictionaries
- 10 Manuscript Description
- 11 Representation of Primary Sources
- 12 Critical Apparatus
- 13 Names, Dates, People, and Places
- 14 Tables, Formulae, Graphics and Notated Music
- 15 Language Corpora
- 16 Linking, Segmentation, and Alignment
- 17 Simple Analytic Mechanisms
- 18 Feature Structures
- 19 Graphs, Networks, and Trees
- 20 Non-hierarchical Structures
- 21 Certainty, Precision, and Responsibility
- 22 Documentation Elements
- 23 Using the TEI

TEI sourcecode

- [Getting and Using the TEI Sources](#)
- [TEI GitHub Repository](#)
- [Bug Reports, Feature Requests, etc.](#)

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

TEI Consortium | [Feedback](#)

Front Matter

Title

- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- ⊕ iv. [About These Guidelines](#)
- ⊕ v. [A Gentle Introduction to XML](#)
- ⊕ vi. [Languages and Character Sets](#)

Back Matter

- ⊕ Appendix A [Model Classes](#)
- ⊕ Appendix B [Attribute Classes](#)
- ⊕ Appendix C [Elements](#)
- ⊕ Appendix D [Attributes](#)
- ⊕ Appendix E [Datatypes and Other Macros](#)
- ⊕ Appendix F [Bibliography](#)
- ⊕ Appendix G [Deprecations](#)
- ⊕ Appendix H [Prefatory Notes](#)
- Appendix I [Colophon](#)

Text Body

- ⊕ 1 [The TEI Infrastructure](#)
- ⊕ 2 [The TEI Header](#)
- ⊕ 3 [Elements Available in All TEI Documents](#)
- ⊕ 4 [Default Text Structure](#)
- ⊕ 5 [Characters, Glyphs, and Writing Modes](#)
- ⊕ 6 [Verse](#)
- ⊕ 7 [Performance Texts](#)
- ⊕ 8 [Transcriptions of Speech](#)
- ⊕ 9 [Dictionaries](#)
- ⊕ 10 [Manuscript Description](#)
- ⊕ 11 [Representation of Primary Sources](#)
- ⊕ 12 [Critical Apparatus](#)
- ⊕ 13 [Names, Dates, People, and Places](#)
- ⊕ 14 [Tables, Formulæ, Graphics and Notated Music](#)
- ⊕ 15 [Language Corpora](#)
- ⊕ 16 [Linking, Segmentation, and Alignment](#)
- ⊕ 17 [Simple Analytic Mechanisms](#)
- ⊕ 18 [Feature Structures](#)
- ⊕ 19 [Graphs, Networks, and Trees](#)
- ⊕ 20 [Non-hierarchical Structures](#)
- ⊕ 21 [Certainty, Precision, and Responsibility](#)
- ⊕ 22 [Documentation Elements](#)
- ⊕ 23 [Using the TEI](#)

Appendix C Elements

1 Appendix C.1 About the Elements Appendix

This appendix gives you links to reference pages for all elements in the TEI Guidelines. There are 573 TEI elements in revision [3c0c64ec4](#) of TEI P5 [Version 3.5.0](#) of the TEI Guidelines.

The elements listed here are in the TEI Namespace: <http://www.tei-c.org/ns/1.0> unless otherwise noted on that element's reference page.

Sorted alphabetically

[a](#) [b](#) [c](#) [d](#) [e](#) [f](#) [g](#) [h](#) [i](#) [j](#) [k](#) [l](#) [m](#) [n](#) [o](#) [p](#) [q](#) [r](#) [s](#) [t](#) [u](#) [v](#) [w](#) [x](#) [z](#) [Show all](#) [Show by module](#)

a

[ab](#) [abbr](#) [abstract](#) [accMat](#) [acquisition](#) [activity](#) [actor](#) [add](#) [additional](#) [additions](#) [addName](#) [address](#) [addrLine](#) [addSpan](#) [adminInfo](#) [affiliation](#) [age](#) [alt](#) [alternate](#) [altGrp](#) [altIdent](#) [altIdentifier](#) [am](#) [analytic](#) [anchor](#) [annotationBlock](#) [anyElement](#) [app](#) [appInfo](#) [application](#) [arc](#) [argument](#) [att](#) [attDef](#) [attList](#) [attRef](#) [author](#) [authority](#) [availability](#)

b

[back](#) [bibl](#) [biblFull](#) [biblScope](#) [biblStruct](#) [bicond](#) [binary](#) [binaryObject](#) [binding](#) [bindingDesc](#) [birth](#) [bloc](#) [body](#) [broadcast](#) [byline](#)

c

[caesura](#) [calendar](#) [calendarDesc](#) [camera](#) [caption](#) [case](#) [castGroup](#) [castItem](#) [castList](#) [catchwords](#) [catDesc](#) [category](#) [catRef](#) [cb](#) [cell](#) [certainty](#) [change](#) [channel](#) [char](#) [charDecl](#) [charName](#) [charProp](#) [choice](#) [cit](#) [citedRange](#) [cl](#) [classCode](#) [classDecl](#) [classes](#) [classRef](#) [classSpec](#) [climate](#) [closer](#) [code](#) [collation](#) [collection](#) [colloc](#) [colophon](#) [cond](#) [condition](#) [constitution](#) [constraint](#) [constraintSpec](#) [content](#) [corr](#) [correction](#) [correspAction](#) [correspContext](#) [correspDesc](#) [country](#) [creation](#) [cRefPattern](#) [custEvent](#) [custodialHist](#)

d

[damage](#) [damageSpan](#) [dataFacet](#) [dataRef](#) [dataSpec](#) [datatype](#) [date](#) [dateline](#) [death](#) [decoDesc](#) [decoNote](#) [def](#) [default](#) [defaultVal](#) [del](#) [delSpan](#) [depth](#) [derivation](#) [desc](#) [dictScrap](#) [dim](#) [dimensions](#) [distinct](#) [distributor](#) [district](#) [div](#) [div1](#) [div2](#) [div3](#) [div4](#) [div5](#) [div6](#) [div7](#) [divGen](#) [docAuthor](#) [docDate](#) [docEdition](#) [docImprint](#) [docTitle](#) [domain](#)

e

[edition](#) [editionStmt](#) [editor](#) [editorialDecl](#) [education](#) [eg](#) [egXML](#) [eLeaf](#) [elementRef](#) [elementSpec](#) [email](#) [emph](#) [empty](#) [encodingDesc](#) [entry](#) [entryFree](#) [epigraph](#) [epilogue](#) [equipment](#) [equiv](#) [eTree](#) [etym](#) [event](#) [ex](#) [exemplum](#) [expan](#) [explicit](#) [extent](#)

f

[facsimile](#) [factuality](#) [faith](#) [fDecl](#) [fDescr](#) [figDesc](#) [figure](#) [fileDesc](#) [filiation](#) [finalRubric](#) [fLib](#) [floatingText](#) [floruit](#) [foliation](#) [foreign](#) [forename](#) [forest](#) [form](#) [formula](#) [front](#) [fs](#) [fsConstraints](#) [fsdDecl](#) [fsDecl](#) [fsDescr](#) [fsdLink](#) [funder](#) [fvLib](#) [fw](#)

g

[gap](#) [gb](#) [gen](#) [genName](#) [geo](#) [geoDecl](#) [geogFeat](#) [geogName](#) [gi](#) [gloss](#) [glyph](#) [glyphName](#) [gram](#) [gramGrp](#) [graph](#) [graphic](#) [group](#)

h

[handDesc](#) [handNote](#) [handNotes](#) [handShift](#) [head](#) [headItem](#) [headLabel](#) [height](#) [heraldry](#) [hi](#) [history](#) [hom](#) [hyph](#) [hyphenation](#)

i

[ident](#) [idno](#) [if](#) [iff](#) [imprimatur](#) [imprint](#) [incident](#) [incipit](#) [index](#) [iNode](#) [institution](#) [interaction](#) [interp](#) [interpGrp](#) [interpretation](#) [item](#) [iType](#)

What can you encode with the TEI?

- The TEI takes a general and agnostic approach to **textual** structure and **verbal** phenomena
 - **textual** is based on a prevalent codex format, but it works for non-codex material, too
 - **verbal** is mostly based on linguistics, but there are general-purpose interpretive tags, too
- In theory, TEI can cope with any texts of any size, language, date, complexity, writing system, or media (in practice--that is another matter)
 - e.g. books, journals, letters, manuscripts, postcards, rolls of papyrus, clay tablets, web pages, gravestones, posters, billboards, etc. and contain any type of text

What a valid TEI document requires

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_lite.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_lite.rng" type="application/xml"
  schematypens="http://purl.oclc.org/dsdl/schematron"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <publicationStmt>
        <p>Publication information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Some text here.</p>
      <figure>
        <graphic url="http://www.tei-c.org/logos/TEI-glow.png"/>
      </figure>
    </body>
  </text>
</TEI>
```

What a valid TEI document requires

1. A root **<TEI>** element that points to the TEI C's XML namespace.
2. TEI header **<teiHeader>**: metadata, which must include:
 - a. **<fileDesc>**: a description of the file, which must include:
 - i. A **<titleStmt>** containing a **<title>**, **<author>**, and any information about others responsible for the document.
 - ii. A **<publicationStmt>** providing publication details about the electronic text in a structured way or as prose inside a **<p>**.
 - iii. A description of the source in a **<sourceDesc>**, which lists bibliographic details about the electronic text's material source in a structured way or in a **<p>**
3. A **<text>** element, the structure of which is represented by elements such as **<front>**, **<body>**, and **<back>**. Within these three structural divisions are further subdivisions often encoded with **<div>** elements.

What a valid TEI document requires

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_lite.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_lite.rng" type="application/xml"
  schematypens="http://purl.oclc.org/dsdl/schematron"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader> ←
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <publicationStmt>
        <p>Publication information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text> ←
    <body>
      <p>Some text here.</p>
      <figure>
        <graphic url="http://www.tei-c.org/logos/TEI-glow.png"/>
      </figure>
    </body>
  </text>
</TEI>
```

TEI and Constraint

The TEI offers around 570 elements and more than 200 attributes.

You cannot read all the books in the library, in other words.

TEI is not a schema. Editors should select TEI elements and attributes from the TEI modules that apply to their projects. TEI offers an unambiguous foundation for representing text as well as standardized mechanisms for customisation.

Good news: few projects require a complete knowledge of the TEI, and **most projects only require about 25 or so elements** in total.

What kind of outputs might you make from TEI?

- What TEI enables you to do is to encode your interpretation of the materials
- What you generate from your TEI encoding is a separate step but ...
- It is important to remember that there is not a one-to-one representation between your TEI and any particular output (like a single view of a digital edition)
- You might generate any number of views of an edition, supplementary indexes, or research aids, introductory editorial material, critical apparatus, metadata lists of people/places/objects/or anything you marked - or you might not generate an edition at all
- You might do this in any number of formats, HTML, MSWord DocX, PDF, etc.
- Or in the necessary format for further analysis in other software (e.g. CSV/Spreadsheets of extracted information)