# Text Encoding Initiative

*Official logo*

The **Text Encoding Initiative** (**TEI**) is a text-centric community of practice in the academic field of digital humanities, operating continuously since the 1980s. The community currently runs a mailing list, meetings and conference series, and maintains an eponymous technical standard, a journal, a wiki, a SourceForge repository and a toolchain.

## 1 TEI guidelines

The *TEI Guidelines*, which collectively define an XML format, are the defining output of the community of practice. The format differs from other well-known open formats for text (such as HTML and OpenDocument) in that it's primarily semantic rather than presentational; the semantics and interpretation of every tag and attribute are specified. Some 500 different textual components and concepts (word,[1] sentence,[2] character,[3] glyph,[4] person,[5] etc.); each is grounded in one or more academic discipline and examples are given.

### 1.1 Technical details

The standard is split into two parts, a discursive textual description with extended examples and discussion and set of tag-by-tag definitions. Schemata in most of the modern formats (DTD, RELAX NG and W3C Schema) are generated automatically from the tag-by-tag definitions. A number of tools support the production of the guidelines and the application of the guidelines to specific projects.

A number of special tags are used to circumvent restrictions imposed by the underlying Unicode; glyph to allow representation of characters that don't qualify for Unicode inclusion[1] and choice to allow overcome the required strict linearity.[6]

Most users of the format do not use the complete range of tags but produce a customisation, using a project-specific subset of the tags and attributes defined by the Guidelines. The TEI defines a sophisticated customization mechanism known as ODD for this purpose. In addition to documenting and describing each TEI tag, an ODD specification specifies its content model and other usage constraints, which may be expressed using schematron.

*TEI Lite* is an example of such a customization. It defines an XML-based file format for exchanging texts. It is a manageable selection from the extensive set of elements available in the full TEI Guidelines.

### 1.2 Examples

The text of the TEI guidelines is rich in examples. There is also a samples page on the TEI wiki[7] which gives examples of real-world projects which expose their underlying TEI.

#### 1.2.1 Prose tags

TEI allows texts to be marked up syntactically at any level of granularity, or mixture of granularities. For example, this paragraph (p) has been marked up into sentences (s) and clauses (cl).[8]

<s> <cl>It was about the beginning of September, 1664, <cl>that I, among the rest of my neighbours, heard in ordinary discourse <cl>that the plague was returned again to Holland; </cl> </cl> </cl> <cl>for it had been very violent there, and particularly at Amsterdam and Rotterdam, in the year 1663, </cl> <cl>whither, <cl>they say,</cl> it was brought, <cl>some said</cl> from Italy, others from the Levant, among some goods <cl>which were brought home by their Turkey fleet;</cl> </cl> <cl>others said it was brought from Candia; others from Cyprus. </cl> </s> <s> <cl>It mattered not <cl>from whence it came;</cl> </cl> <cl>but all agreed

<cl>it was come into Holland again.</cl> </cl> </s>

### 1.2.2 Verse

TEI has tags for marking up verse. This example (taken from the French translation of the TEI Guidelines) shows a sonnet[9]

<div type="sonnet"> <lg type="quatrain"> <l>Les amoureux fervents et les savants austères</l> <l> Aiment également, dans leur mûre saison,</l> <l> Les chats puissants et doux, orgueil de la maison,</l> <l> Qui comme eux sont frileux et comme eux sédentaires.</l> </lg> <lg type="quatrain"> <l>Amis de la science et de la volupté</l> <l> Ils cherchent le silence et l'horreur des ténèbres ;</l> <l> L'Érèbe les eût pris pour ses coursiers funèbres,</l> <l> S'ils pouvaient au servage incliner leur fierté.</l> </lg> <lg type="tercet"> <l>Ils prennent en songeant les nobles attitudes</l> <l>Des grands sphinx allongés au fond des solitudes,</l> <l>Qui semblent s'endormir dans un rêve sans fin ;</l> </lg> <lg type="tercet"> <l>Leurs reins féconds sont pleins d'étincelles magiques,</l> <l> Et des parcelles d'or, ainsi qu'un sable fin,</l> <l>Étoilent vaguement leurs prunelles mystiques.</l> </lg> </div>

### 1.2.3 Choice tag

The choice tag is used to represent sections of text which might be encoded or tagged in more than one possible way. In the following example, based on one in the standard, choice is used twice, once to indicate an original and a corrected year and once to indicate an original and regularised spelling.[10]

<p xml:id=\char"0022\relax{}p23">Lastly, That, upon his solemn oath to observe all the above articles, the said man-mountain shall have a daily allowance of meat and drink sufficient for the support of <choice> <sic>1724</sic> <corr>1728</corr> </choice> of our subjects, with free access to our royal person, and other marks of our <choice> <orig>favour</orig> <reg>favor</reg> </choice>.

## 2 ODD

**One Document Does it all** ("ODD") is a literate programming language for XML schemas.[11][12][13][14]

In literate-programming style, ODD documents combine human-readable documentation and machine-readable models using the Documentation Elements module of the Text Encoding Initiative. Tools generate localised and internationalised HTML, ePub, or PDF human-readable output and DTDs, W3C XML Schema, Relax NG Compact Syntax, or Relax NG XML Syntax machine-readable output.

The Roma web application[15] is built around the ODD format and can use it to generate schemas in DTD, W3C XML Schema, Relax NG Compact Syntax, or Relax NG XML Syntax formats, as used by many XML validation tools and services.

ODD is the format used internally by the Text Encoding Initiative for their eponymous technical standard.[16] Although ODD files generally describe the difference between a customized XML format and the full TEI model, ODD also can be used to describe XML formats that are entirely separate from the TEI. One example of this is the W3C's Internationalization Tag Set which uses the ODD format to generate schemas and document its vocabulary.[17][18]

## 3 TEI customizations

TEI customizations are specializations of the TEI XML specification for use in particular fields or by specific communities.

- EpiDoc (Epigraphic Documents)
- Charters Encoding Initiative
- Medieval Nordic Text Archive (Menota)

Customization in the TEI is done through the ODD mechanism mentioned above. In truth since its P5 version, all so-called 'TEI Conformant' uses of the TEI Guidelines are based on a TEI customization documented in a TEI ODD file. Even when users choose one of the off-the-shelf pre-generated schemas to validate against, these have been created from freely available customization files.

## 4 Projects

The format is used by many projects worldwide. Practically all projects are associated with one or more universities. Some well-known projects that encode texts using TEI include:

## 5 History

Prior to the creation of TEI, humanities scholars had no common standards for encoding electronic texts in a manner which would serve their academic goals (Hockey 1993, p. 41). In 1987, a group of scholars representing fields in humanities, linguistics, and computing convened at Vassar College to put forth a set of guidelines known as the "Poughkeepsie Principles". These guidelines directed the development of the first TEI standard, "P1"[19][20]

- **1987** Work on what would become the TEI started by the Association for Computers and the Humanities,[21] the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing.[22] This culminated in the *Closing statement of the Vassar Planning Conference*[23]
- **1994** TEI P3 released[24] co-edited by Lou Burnard (at Oxford University) and Michael Sperberg-McQueen (then at the University of Illinois at Chicago, later at the W3C).
- **1999** TEI P3 updated.
- **2002** TEI P4 released, moving from SGML to XML; adoption of Unicode, which XML parsers are required to support.[25]
- **2007** TEI P5 released, including integration with the xml:lang and xml:id attributes from the W3C[26] (these had previously been attributes in the TEI namespace), regularization of local pointing attributes to use the hash (as used in HTML) and unification of the ptr and xptr tags. Together these changes with many more new additions make P5 more regular and bring it closer to current xml practice as promoted by the W3C and as used by other XML variants. Maintenance and feature update versions of TEI P5 have been released at least twice a year since 2007.
- **2011** TEI P5 v2.0.1 released with support for Genetic editing.[27] (among many other additions the Genetic editing features allow encoding of texts without interpretation as to their specific semantics.)

## 6 References

[1] "Element w (word) - TEI P5".

[2] "Element s (s-unit) - TEI P5".

[3] "Element c (character) - TEI P5".

[4] "Element g (character or glyph) - TEI P5".

[5] "Element person (person) - TEI P5".

[6] "Element choice - TEI P5".

[7] "Samples of TEI texts". *wiki.tei-c.org*. 2011. Retrieved 17 April 2012.

[8] "17 Simple Analytic Mechanisms - TEI P5: — Guidelines for Electronic Text Encoding and Interchange". *tei-c.org*. 2012. Retrieved 15 April 2012.

[9] "TEI element lg (groupe de vers)". *tei-c.org*. 2012. Retrieved 15 April 2012.

[10] "TEI element choice". *tei-c.org*. 2012. Retrieved 15 April 2012.

[11] Bauman, Syd; Flanders, Julia (2004), "ODD customizations", *Extreme Markup Languages 2004*.

[12] Burnard, Lou; Rahtz, Sebastian (2004), "RelaxNG with Son of ODD", *Extreme Markup Languages 2004*.

[13] Reiss, Kevin M. (2007), *Literate Documentation for XML* (PDF), Urbana-Champaign, Illinois: Digital Humanities 2007.

[14] Burnard, Lou; Rahtz, Sebastian (June 2013). "A complete schema definition language for the Text Encoding Initiative". *XML London 2013*: 152–161. doi:10.14337/XMLLondon13.Rahtz01. ISBN 978-0-9926471-0-0.

[15] Roma web application.

[16] Burnard, Lou; Bauman, Syd, eds. (2007), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Charlottesville, Virginia, USA: TEI Consortium.

[17] W3C ITS and TEI ODD file.

[18] Savourel, Yves; Kosek, Jirka; Ishida, Richard, eds. (2008), "5.2 ITS and TEI", *Best Practices for XML Internationalization*, W3C Working Group.

[19] Ahronheim, J.R. (1998). "Descriptive metadata: Emerging standards.". *Journal of Academic Librarianship*. **24** (5): 395.

[20] Cantara, L. (2005). "The text-encoding initiative: Part 1". *OCLC Systems & Services*. **21** (1): 36–39. doi:10.1108/10650750510578136.

[21] ach.org

[22] "Historical background", section iv.2 of TEI P5: Guidelines for Electronic Text Encoding and Interchange.

[23] "Closing statement of the Vassar Planning Conference". *tei-c.org*. 2009. Retrieved 15 April 2012.

[24] "TEI Guidelines". Retrieved 2010-06-18.

[25] "2", *XML Basics*, retrieved 2011-07-09

[26] "Extensible Markup Language (XML) 1.0 (Fifth Edition)". *w3.org*.

[27] "P5 version 2.0.1 release notes". *tei-c.org*. 2012. Retrieved 15 April 2012.

## 7 External links

- TEI Consortium Web site with a list of TEI projects, a form for adding your project and wiki
- Journal of the TEI

- TEI Lite: An Introduction to Text Encoding for Interchange
- TEI @ Oxford (hosted at Oxford University) with development and backup versions of much of the core content.
- TEI development (hosted at SourceForge.net) with bugtracker, version control, etc.
- Larger list of TEI Projects
- What is the TEI? (Introductory overview by Lou Burnard)

# 8 Text and image sources, contributors, and licenses

## 8.1 Text

- **Text Encoding Initiative** *Source:* https://en.wikipedia.org/wiki/Text_Encoding_Initiative?oldid=771113633 *Contributors:* Nealmcb, Cherkash, Peak, Rursus, (:Julien:), Lucky 6.9, Bender235, CanisRufus, A2Kafir, Schnolle, Wtshymanski, BDD, Jcummings, Stuartyeates, Woohookitty, Kesla, Koavf, Margosbot~enwiki, Sderose, Geekgrrrrrrl, Dsewell, Benlisquare, Hede2000, Gabrielbodard, Wujastyk, H@r@ld, Mhkay, Jeremy Butler, SmackBot, ARK, Thumperward, Chris55, Cydebot, Thijs!bot, Mr pand, Shaunm, Craigsapp, A3nm, David Eppstein, Girl2k, Wikidemon, Andy Dingley, Oskilla, M4gnum0n, Ducloy, Addbot, Ghettoblaster, MrOllie, Lightbot, Xqbot, Control.valve, SebastianHellmann, Farnhamian, FrescoBot, Craddoke, Redrose64, Jonesey95, Σ, Lotje, XPtr, Sallyrenee, Rmahfoud, Dewritech, Digitize it, Rettinghaus, BG19bot, Vilius Normantas, Johannes.daxenberger, Martinholmes, BattyBot, ChrisGualtieri, AndiPersti, SeMelmoth, Nkultra, Klnasy, Monkbot, KasparBot, SkaakBurger, Exemplo347, Skirden and Anonymous: 30

## 8.2 Images

- **File:Text_Encoding_Initiative_TEI-800.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/5/5b/Text_Encoding_Initiative_TEI-800.jpg *License:* CC BY 3.0 *Contributors:* http://www.tei-c.org/About/Logos/ *Original artist:* Andrew Paulin and Rowan Wilson

## 8.3 Content license

-