Attribution-NonCommercial-ShareAlike 2.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

**Attribution.** You must give the original author credit.

**Noncommercial.** You may not use this work for commercial purposes.

**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

[Disclaimer](#)
The Text Encoding Initiative
An Introduction

Dr Susan Schreibman
University of Maryland
May 2003

The TEI
• history of the TEI
• philosophy behind TEI
• practicalities of using it

The Text Encoding Initiative
developed between 1987-1994
• joint industry/educational/govt/non-profit initiative/w hundreds of participants
• a subset of SGML developed specifically for humanities applications
• TEI P1 published in 1990
• TEI P2 published between 1992-93
• TEI P3 published in 1994
• TEI P4 published in 1999
• TEI U5 published in 1995 (TEI Lite)
• TEI P4 (XML) published in 2002

An International Consortium:
old website
http://www.uic.edu/orgs/tei/
• funders
  – the Natl Endowment of the Humanities
  – Commission of the EU
  – Andrew W Mellon Foundation
  – Social Science & Humanities Research Council of Canada
• sponsors
  – Assoc of Computers and the Humanities
  – Assoc for Computational Linguistics
  – Assoc for Literary and Linguistic Computing

New Consortium to maintain the TEI formed in 1999
• hosted by:
  – University of Virginia
  – Brown University
  – Oxford University
  – University of Bergen
• more info at:
  – http://www.tei-c.org/

the Guidelines provide a means of making explicit certain features of a text in such a way as to aid the processing of that text by computer programs running on different machines. This process of making explicit we call *markup* or *encoding*.

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div1-c15b2b3b3b2.html
The guidelines create a standard and guide for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally.

**Intended Use**
- guidance for individual or local practice in text creation and data capture
- support for data interchange
- support of application-independent processing

**Goals of the TEI (1987):**
- suffice to represent the textual features needed for research;
- be simple, clear, and concrete;
- be easy for researchers to use without special-purpose software;
- allow the rigorous definition and efficient processing of texts;
- provide for user-defined extensions;
- conform to existing and emergent standards.

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div1-c15b2b3b3b2.html
Entity Sets

- additional specialist features be easy to add to (or remove from) a text;

- multiple parallel encodings of the same feature should be possible;

- the richness of markup should be user-defined, with a very small minimal requirement;

- adequate documentation of the text and its encoding should be provided;

- a DTD flexible enough to encompass a wide range of texts, periods & purposes
  - history
  - literature
  - art history
  - linguistics
  - ancient texts
  - medieval texts
  - modern texts
has become the de facto standard
- most major humanities computing projects utilise it
- in theory, allows for the exchange of texts across projects and archives
- helps to ensure uniform encoding of text: extremely important for both humans and parsers

What does TEI facilitate?
- **structural divisions within a text**
  - title-page, chapter, scene, stanza, line, etc
- **typographical elements**
  - changes in typeface, special characters, etc
- **other textual features**
  - grammatical structure, location of illustrations, variant forms, etc

facilitates the long-term preservation of electronic texts
- repositories have one basic DTD to preserve and understand
  - though that DTD may have many "flavors"

all TEI documents follow the same essential format
- **TEI header**
  - documents the electronic edition being created
- **TEI body**
  - contains the content being created

TEI Header

TEI Body
TEI flavors: The TEI Pizza Site

The Full TEI Tag Set

- developed to meet a range of encoding needs
  - across disciplines
  - across language needs
    - literature vs linguistics
    - developed in a pre-internet environment

Concept behind Pizza

“The TEI Guidelines define several hundred SGML elements and associated attributes, which can be combined to make many different DTDs, suitable for many different purposes, either simple or complex. With the aid of the Pizza Chef, you can build a DTD that contains just the elements you want, suitable for use with any SGML or XML processing system.”

Pizza concept

- there are various base tag sets (including)
  - prose
  - verse
  - drama
  - dictionaries
  - speech
  - mixed base

*the common core of textual features be easily shared;*

Pizza concept

- individuals choose various "toppings", ie additional tag sets based on encoding needs (including)
  - linking
  - figures
  - analysis (linguistic analysis)
  - transcription
  - textcrit
  - names.dates
  - certainty
**additional specialist features be easy to add to (or remove from) a text;**

---

**Step 2: choose your toppings**

Which toppings you choose above, you will always get the core ingredients, defining common core elements and the TEI Entities. These are enough for you, provided also, that you can choose or hire as easy as you want the following additional toppings. (The default definitions are those included by the TEI Lite 1.0):

<table>
<thead>
<tr>
<th>Topping</th>
<th>Additional Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eating</td>
<td>Additional features for hyphenation, segmentation, and alignment</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for encoding idioms, proverbs, and formulae</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for determining and indexing linguistic analysis</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for annotating and indexing acoustically</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of monetary currencies (e.g. amount, date)</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of musical scores (e.g. timing, note)</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of scientific and technical terms</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of astronomical and meteorological terms</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of geographical and historical terms</td>
</tr>
<tr>
<td>Eating</td>
<td>Additional features for the transcription of organizational and institutional terms</td>
</tr>
</tbody>
</table>

---

**Entity Sets**

**additional specialist features be easy to add to (or remove from) a text;**

---

**TEI Lite**

- a subset of of TEI designed to meet average encoding needs
- no choosing of bases and toppings
- no choosing of entity sets – all is chosen for you
- "conceived of as a simple demonstration of how the TEI encoding scheme might be adopted to meet 90% of the needs of 90% of the TEI user community."

http://www.hcu.ox.ac.uk/TEI/Lite/teiu5-div-c15b2b3b1b3.html

---

**Goals of TEI Lite**

- it should include most of the TEI “core” tag set, since this contains elements relevant to virtually all text types and all kinds of text-processing work;
- it should be able to handle adequately a reasonably wide variety of texts, at the level of detail found in existing practice (for ex the Oxford Text Archive);
- it should be useful for the production of new documents as well as encoding of existing ones;

---

**TEI Lite is most suited to**

- printed texts
- structural encoding
- light content encoding
- electronic repositories which want to make lightly encoded texts available to scholars

---

**how does it fulfill its goals?**

- it includes most features of the base prose tag set;
- but skips more specialised usages, for ex:
  - TEI Lite includes `<name>`
- if one chooses the “names and dates” additional tag set one gets:
• `<persName>` contains a proper noun or proper-noun phrase referring to a person, possibly including any or all of the person's forename, surname, honorific, added names, etc.
• `<surname>` contains a family (inherited) name, as opposed to a given, baptismal, or nick name.
• `<forename>` contains a forename, given or baptismal name.
• `<roleName>` contains a name component which indicates that the referent has a particular role or position in society, such as an official title or rank.
• `<addName>` contains an additional name component, such as a nickname, epithet, or alias, or any other descriptive phrase used within a personal name.
• `<nameLink>` contains a connecting phrase or link used within a name but not regarded as part of it, such as ‘van der’ or ‘of’.
• `<genName>` contains a name component used to indicating generational information, such as ‘Junior’, or a number used in a monarch's name.

how does TEI Lite fulfill its goal?

• for encoding poetry, TEI Lite offers:
  – `<lg>` for encoding line groups (stanzas)
  – `<l>` for encoding a line of poetry

the verse tag set offers:

• “numbered” `<lg>` elements are provided, by analogy with the `divn` class elements
• a special purpose `<caesura>` element is provided, to allow for segmentation of the verse line
• a set of attributes is provided for the encoding of rhyme scheme and metrical information

from Pope’s Essay on Criticism

```xml
<text><front>... </front>
<body met='-+|-+|-+|-+|-+/' rhyme='aa'>
<lg n=1 type="verse paragraph">
<l>‘Tis hard to say, if greater Want of Skill<br/>
Appear in <hi>Writing</> or in <hi>Judging</> ill;</l>
<l>But, of the two, less dang’rous is th’Offence;</l>
<l>To tire our <hi>Patience</> than mis-lead<br/>
our <hi>Sense</>;</l>
</lg></body>
</text>
```

caesura element

```xml
<l>In a somer seson, <caesura> whan softe was the sonne, </l>
<l>I shoop me into shroudes <caesura> as I a sheep were, </l>
<l>In habite as an heremite <caesura> unholy of werkes, </l>
<l>Went wide in this world <caesura> wondres to here. </l>
```
attributes for encoding rhyme schemes and metrical information

- **met** contains a user-specified encoding for the conventional metrical structure of the element.
- **real** contains a user-specified encoding for the actual realization of the conventional metrical structure applicable to the element.
- **rhyme** specifies the rhyme scheme applicable to a group of verse lines.

Why choose TEI or TEI Lite?

- is your purpose to do light encoding of texts – structural information (paras, lg, l, etc)
- do you need header information only? – an image gallery
- what is the purpose of detailed metrical markup? – (not a rhetorical question!)
- what are the breadth vs depth issues in archive encoding?
  - (also not a rhetorical question)

```xml
<lg type=stanza>
  <lg type=sestet>
    <l>In the first year of Freedom's second dawn
    <l>Die George the Third; although no tyrant, one
    <l>Who shielded tyrants, till each sense withdrawn
    <l>Left him nor mental nor external sun:
    <l>A better farmer ne'er brushed dew from lawn,
    <l>A worse king never left a realm undone!
  </lg>
</lg>
```
```xml
<lg type=couplet>
  <l>He died &mdash; but left his subjects still behind,
  <l>One half as mad &mdash; and t'other no less blind.
</lg>
```