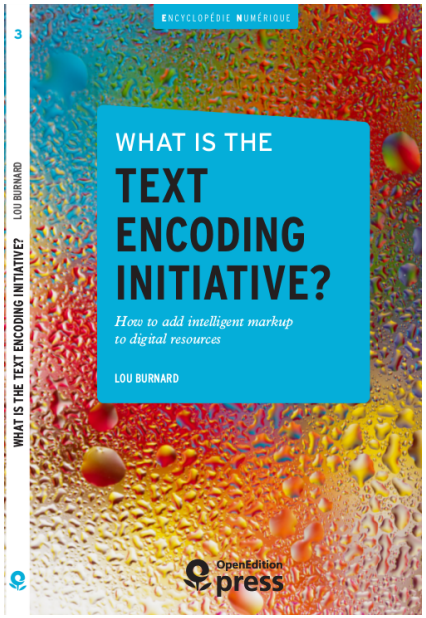
What is the Text Encoding Initiative? (and how did it get to be that way?)

Lou Burnard Consulting

What is the TEI?



- an organization or an institution?
- a club or a religion or a fashion?
- a technical specification, or a framework for making one?
- a community... a way of thinking... a shared perception

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called "Mir" ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called 'mainframes'

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called "Mir" ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called 'mainframes'

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called "Mir" ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called 'mainframes'

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called ”Mir” ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called ‘mainframes’

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called ”Mir” ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called ‘mainframes’

Do you remember 1987 ?

The Text Encoding Initiative was born into a very different world...

- The world wide web did not exist
- The channel tunnel was under construction (at last)
- A state known as the Soviet Union launched a space station called ”Mir” ... and suffered a terrible disaster at Chernobyl
- Records of the year: Raising Hell (Run DMC) and Graceland (Paul Simon)
- and any serious computing was done on what we called ‘mainframes’

... though some of it looks quite familiar ...

- Such disciplines as ‘corpus linguistics’ and ‘artificial intelligence’ were already demonstrating the need to work with large-scale digital textual resources
- Text processing was a recognised field, with pioneering developments both in lexicography (OED), and in document processing systems (TeX, Scribe, tRoff..)
- The Internet existed (though only in academia) and there was much noise about ‘hypertextuality’
- Two familiar technical challenges were already evident : data preservation; data compatibility as new technologies such as CD emerged

Origins of the Text Encoding Initiative

- Spring 1987: European workshops on standardisation of historical data (J.P. Genet, M Thaller)
- Autumn 1987: NEH funds an exploratory international workshop on the feasibility of defining "text encoding guidelines"

Vassar College, Poughkeepsie

Today's question:

- So the TEI is *very old*!
- It comes from a time before the Web, before the DVD, the mobile phone, cable tv, or Microsoft Word
- Not much in computing survives 5 years, never mind 20
- What relevance can it possibly have today?
- Why is it still here, and how has it survived?

The TEI mission

Recognising the demotic potential of the digital the TEI («Text Encoding for Interchange») defined its mission as follows :

- to facilitate thecreation, exchange, and integration of textual data in digital form
  - every kind of text
  - in every language
  - for any purpose, from any culture
- The TEI recommendations are intended for ...
  - beginners, seeking well-established solutions to well-understood problems
  - experts, seeking to create new solutions

Its original design goals

- provide recommendations derived from the existing consensus, where this could be determined
- prefers general solutions to discipline-specific ones
- supporting both specialisation and extension

The TEI was not designed to provide a complete answer 'out of the box'

Why was this considered necessary ?

- The rise and rise of mutually incompatible data formats, hand in hand with the evolution of new technologies !
- And also perhaps a desire to bring traditional philology up to date

A TEI time line

1988 - 1994	Development undertaken by international research project, with funding from US and EU: versions TEI P1 (1990), P2 (1992), P3 (1994)
1995 - 1999	Promotion and take-up of TEI (unfunded)
2000	Establishment of TEI Consortium, incorporated 30 Dec
2001 - 2003	XML conversion of P3 as TEI P4; Council oversees production of complete revision as TEI P5
2003 -	Regularly updated releases of TEI P5; 36 releases since 2005; latest version 3.0.0 April 2016

1988 : a period of transition

- 'Humanities Computing' was beginning to invent itself, as a kind of "interdisciplinary service"
- a dialogue between specialists in computer science and in the humanities was beginning
- some computing centres saw the potential of research activities as a means of enhancing their services
- some research centres saw the potential of computing expertise as a means of enhancing their research.

A fruitful synergy between researcher and engineer...

The Poughkeepsie Principles

Closing Statement of Vassar Conference
The Preparation of Text Encoding Guidelines

Poughkeepsie, New York
13 November 1987

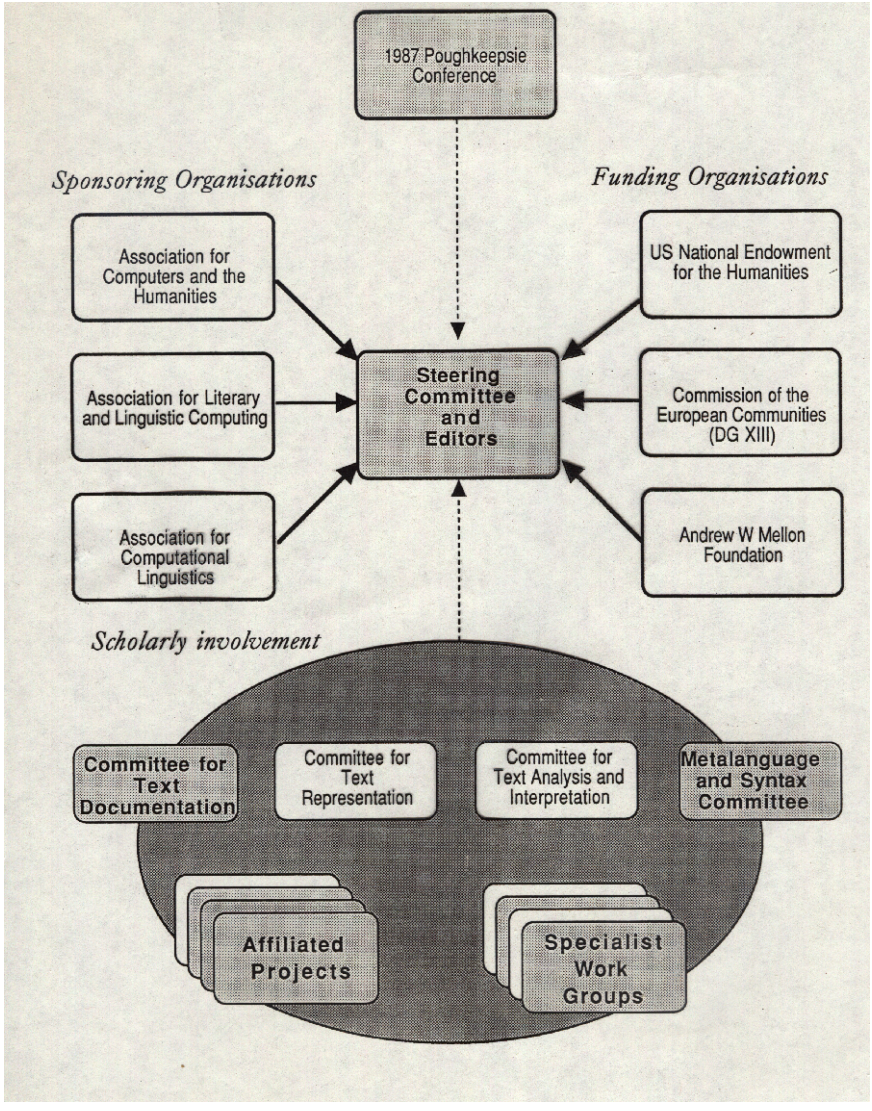
1. The guidelines are intended to provide a standard format for data interchange in humanities research.
2. The guidelines are also intended to suggest principles for the encoding of texts in the same format.
3. The guidelines should
    1. define a recommended syntax for the format,
    2. define a metalanguage for the description of text-encoding schemes,
    3. describe the new format and representative existing schemes both in that metalanguage and in prose.
4. The guidelines should propose sets of coding conventions suited for various applications.
5. The guidelines should include a minimal set of conventions for encoding new texts in the format.
6. The guidelines are to be drafted by committees on
    1. text documentation
    2. text representation
    3. text interpretation and analysis
    4. metalanguage definition and description of existing and proposed schemes,

    coordinated by a steering committee of representatives of the principal sponsoring organizations.
7. Compatibility with existing standards will be maintained as far as possible.
8. A number of large text archives have agreed in principle to support the guidelines in their function as an interchange format. We encourage funding agencies to support development of tools to facilitate this interchange.
9. Conversion of existing machine-readable texts to the new format involves the translation of their conventions into the syntax of the new format. No requirements will be made for the addition of information not already coded in the texts.

The principles agreed upon at the Poughkeepsie Planning Conference are expounded in more detail and supplemented with other material in the sections which follow.

http://www.tei-c.org/Vault/ED/edp01.htm

TEI organization (1991)



The ground work of the TEI was undertaken by four 'working committees' and two 'editors':

- Documentation : composed of experts in bibliography and data archives
- Metalanguage : composed of computer scientists
- Text Analysis and Interpretation : composed of theoretical linguists
- Text Representation : ... everyone else

Analysis vs. representation

Knocking their heads together

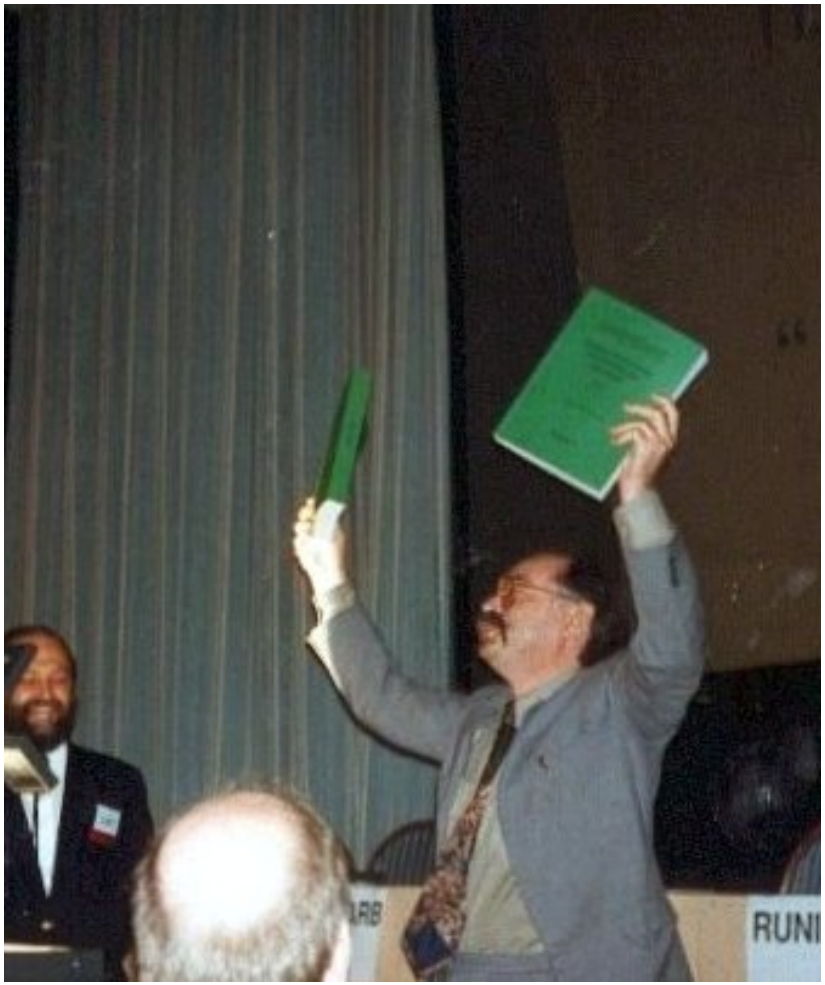
The work of these commitees often overlapped. The two TEI editors had the job of applying *Ockham's razor* to their outputs, as vigorously as possible...

Nevertheless, the TEI still proposes multiple ways of representing (for example)

- linguistic segmentation
- interpretative encoded annotations
- documentation of interpretive codes
- in line or standoff markup
- ...

(This is another reason for not using TEI All!)

1994 : P3



- April 1994: TEI (P3) announced at ALLC-ACH conference Paris
- May: the 'Green Books' appear at SGML94, Montreux.
- Dec 1994, first 'TEI Metaworkshop' held, in Chicago.

1994-1999

The adoption of the TEI and the influence of its ideas are difficult to trace, because it so rapidly became an invisible part of the research infrastructure. Some key events :

- In 1996, Michael Sperberg McQueen, principal editor of the TEI, was appointed co-editor of the emerging W3C standard XML
- In 1997, the 10th anniversary of the TEI was marked by an international conference at Brown University
- In 1998, the US Digital Libraries Federation (DLF) organized a meeting at Washington to discuss the possibility of updating TEI from SGML to XML
- In 1999 a second edition of P3 appeared with some revisions and corrections, and one addition (the <ab> element)
- 1999-2001 : MASTER project develops TEI-conformant manuscript description module

Who owns the output of an international collaborative research project? Who has the right and duty to maintain it?

2000 : Birth of the TEI Consortium

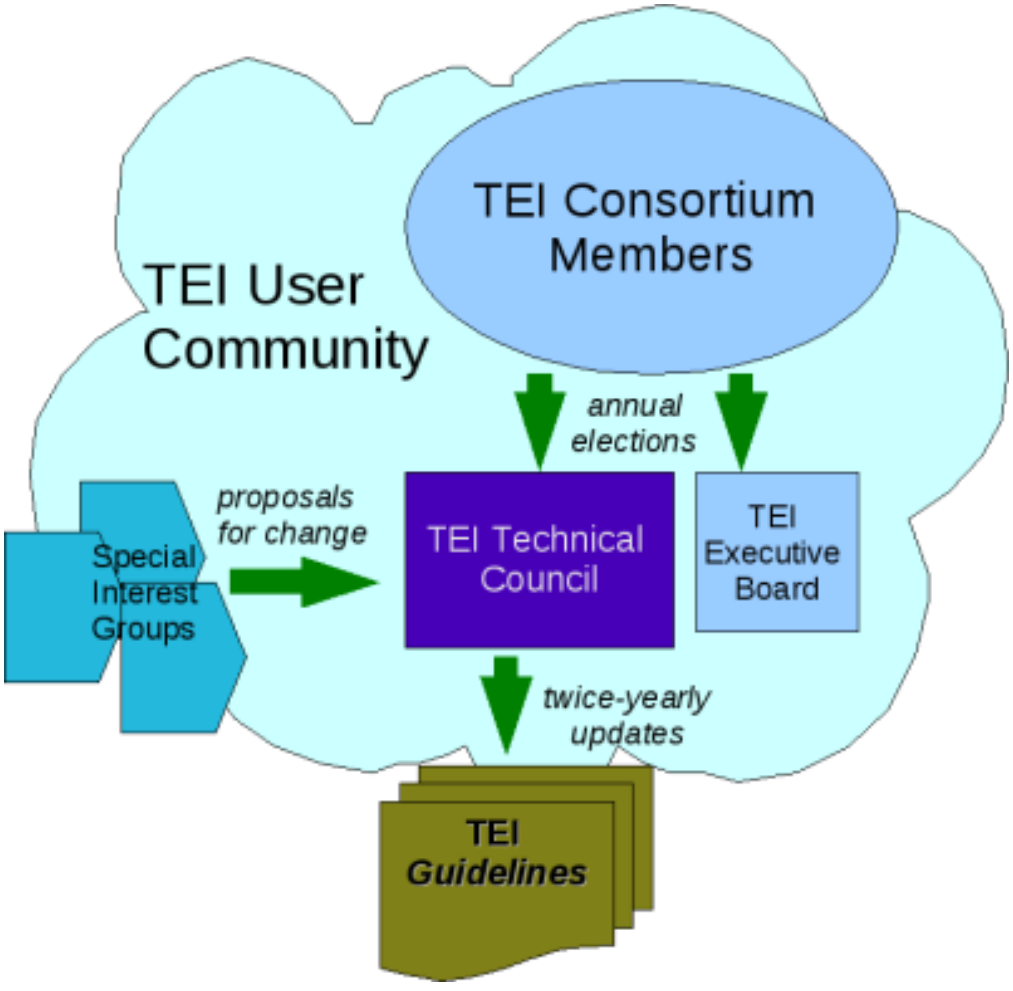
After much work by some key users of the TEI (notably DH centres in London, Virginia, Brown, Oxford, and Bergen) the TEI Consortium was incorporated as a not-for-profit membership association in December 2000.

Goals of the consortium (apart from establishing its own existence)

- ensure the maintenance of the TEI system
- implement the most urgent revisions:
  - an XML version
  - expansion of the subjects treated in line with community needs
- definition of a business model able to sustain future development and maintenance of the TEI community's scientific efforts

http://www.ariadne.ac.uk/issue24/tei/

TEI organization (today)



The TEI is no longer a research project

- A community-driven effort
- Development and maintenance is driven by the Technical Council
- Elected membership
- Responsible to paying members of the Consortium
- Participation for both personal and institutional members

What's *not* TEI?

Originally, the TEI ruled out of consideration...

- the web (it didn't yet exist)
- text formatting or processing (tex, scribe...)
- digital images (without transcription)
- encoding of facts or objects (database country)
- software

its original focus (metadata, text, textual and linguistic analysis) has now expanded to include all of the above

The current TEI landscape

- Basic organization of continuous texts
- Diplomatic transcriptions, images, multimedia, annotations...
- Formal data : dates, names and named entities (people, places, organizations, objects)
- Paratextual data, documentary evidence, textual transmission
- Linguistic analyses of all kinds (including speech and music)
- Documentation and generation of encoding systems
- Et cetera: see http://www.tei-c.org/P5/Guidelines/

... in short, a new kind of Encyclopaedia!

The TEI Architecture

- TEI is a *modular* system. You use it to create an encoding system that reflects your own needs, by choosing from the TEI's pre-defined *modules*
- Each module defines a set of elements and their attributes
- you can choose just the elements you want, and also (within limits) change their properties
- you can add in non-TEI elements, either from other standards or completely new

Define your goals clearly *before* trying to use the TEI !

TEI modules

Name	P5 chapter
analysis	Simple Analytic Mechanisms
certainty	Certainty and Responsibility
core	Elements Available in All TEI Documents
corpus	Language Corpora
dictionaries	Dictionaries
drama	Performance Texts
figures	Tables, Formulae, and Graphics
gaiji	Representation of Non-standard Characters and Glyphs
header	The TEI Header
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcriptions of Speech
tagdocs	Documentation Éléments
tei	The TEI Infrastructure
textcrit	Critical Apparatus
textstructure	Default Text Structure
transcr	Representation of Primary Sources
verse	Verse

How do you choose?

- Take the lot ! (not a very smart choice)
- Use a predefined selection (TEI Lite, TEI Bare...)
- Roll your own – according to your specific project needs

Yes, this means that you have to know about all the possibilities ...

Roma an online tool for helping in this task

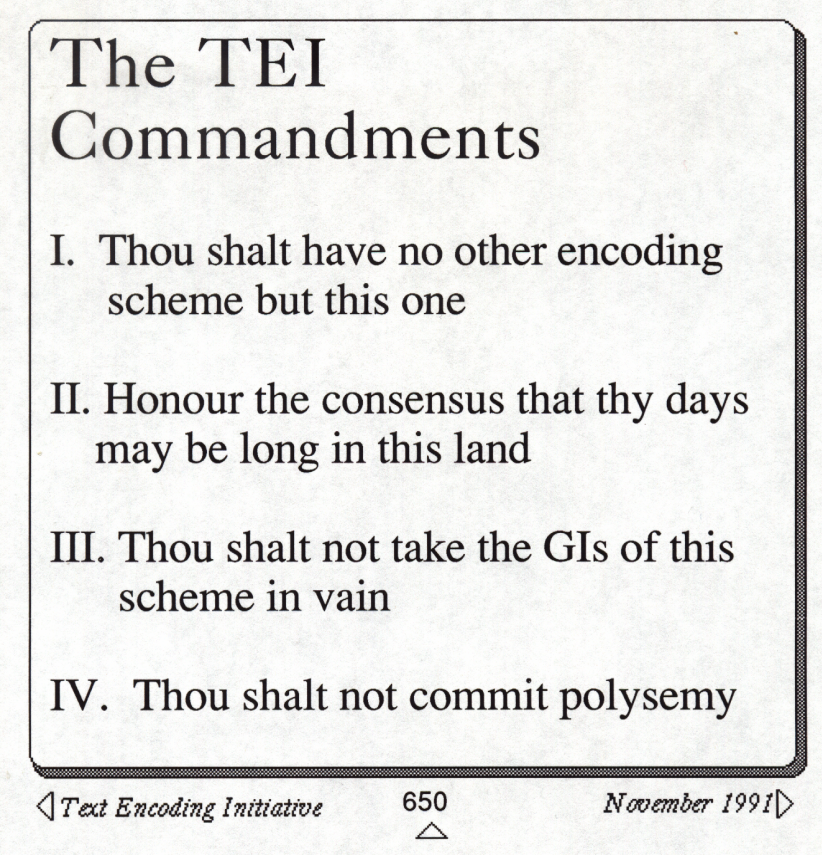
http://www.tei-c.org/Roma/

... leading to the many flavours of TEI

Suppose you want to encode a bibliographic description. The TEI makes you choose between :

- <bibl> (contains any combination of bibliographic elements and text, or just text)
- <biblStruct> or <biblFull> (both containing a specific set of bibliographic elements in a specific order and no text)
- (and other things...)

But surely, a standard exists for the sake of conformance, right?



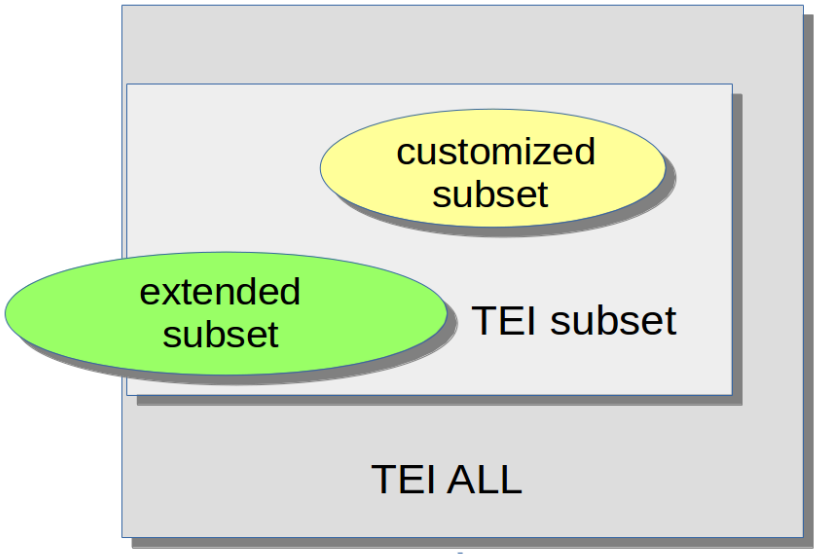
The TEI spirit

So what does it mean to be conformant to the TEI ?

- tagging by consensus
- using a defined lexicon
- respect for diversity

Standardisation should not mean «do what I do» but rather «explain your decisions in terms I can understand »

The TEI was designed to support many approaches



An ODD can...

- just make a selection from TEI All (TEI subset)
- combine that with some additional constraints (customized subset)
- add new components (extended subset)

What does it mean to be 'TEI Conformant'?

- **be honest**: every XML element claiming to belong to the TEI namespace should respect the semantics of its TEI definition
- **be explicit**: the formal documentation provided by a TEI ODD makes make explicit how you have chosen to use the TEI
- The TEI elements in a conformant TEI document must be valid with respect to the TEI All schema

The purpose of these rules is to make 'blind interchange' easier... though they cannot guarantee it

Why is the TEI still of interest?

- you could markup your text in HTML 5, or design your own tagging system
- (but what a waste of time that would be)
- you could use a database system
- (but you would need some nontrivial technical support)

Understanding the TEI requires you to master a 'sweet spot', midway between technology and the humanities

Why is the TEI still of interest?

There are two reasons why standards fail :

- they are based on an immature theory
- ”not invented here”: the user community is fragmented or diverse

How does one test a theory ?

A TEI customisation can :

- control the possible values of attributes
- apply additional semantic or other constraints on element content (eg co-dependency)
- remove elements from a schema
- add new elements in different namespaces

These mechanisms make it easy to test new ideas while remaining TEI conformant.

Not Invented Here?

- TEI P5 includes many I18N features ...
- Like other XML schemas, TEI is hospitable to other namespaces
- A TEI schema can interoperate with other XML standards:
  - SVG for graphics
  - MathML for maths
  - MEI for music
  - ....
- A TEI element definition may specify its equivalent in some other ontology using the <equiv>element
- But fundamentally, the TEI's view of 'what text really is' remains a familiar and traditional one

Darwinism works ...

- Customize the TEI using your own namespace
- Document your changes with an ODD
- Discuss your revisions on the TEI-L list; or form a SIG !
- Propose useful corrections and modifications to the Council, using for example a ”feature request” on http://tei.sf.net
- There’s a new version of TEI P5 roughly twice a year ...

Oh, and don’t forget to join the Consortium!

Some vital links

- http://www.tei-c.org
- http://github.com/TEIC/TEI
- http://listserv.brown.edu/archives/cgi-bin/wa?SUBED1=tei-l&A=1