Introduction to the TEI

TEI@Oxford

February 2010

An Introduction to the TEI

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts chiefly in the humanities, social sciences and linguistics.

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

The TEI aims to be independent of schema language

The TEI encoding scheme is a framework providing:

- definitions and names for several hundred useful textual distinctions
- a set of modules that can be used to generate schemas making those distinctions
- a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
- a very simple consensus-based way of organizing and structuring textual (and others) resources...
- ... which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats

Relevance of the TEI

Why would you want those things?

- because we need to interchange resources
  - between people
  - (increasingly) between machines
- because we need to integrate resources
  - of different media types
  - from different technical contexts
- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data

Relevance of the TEI

Why would you want those things?

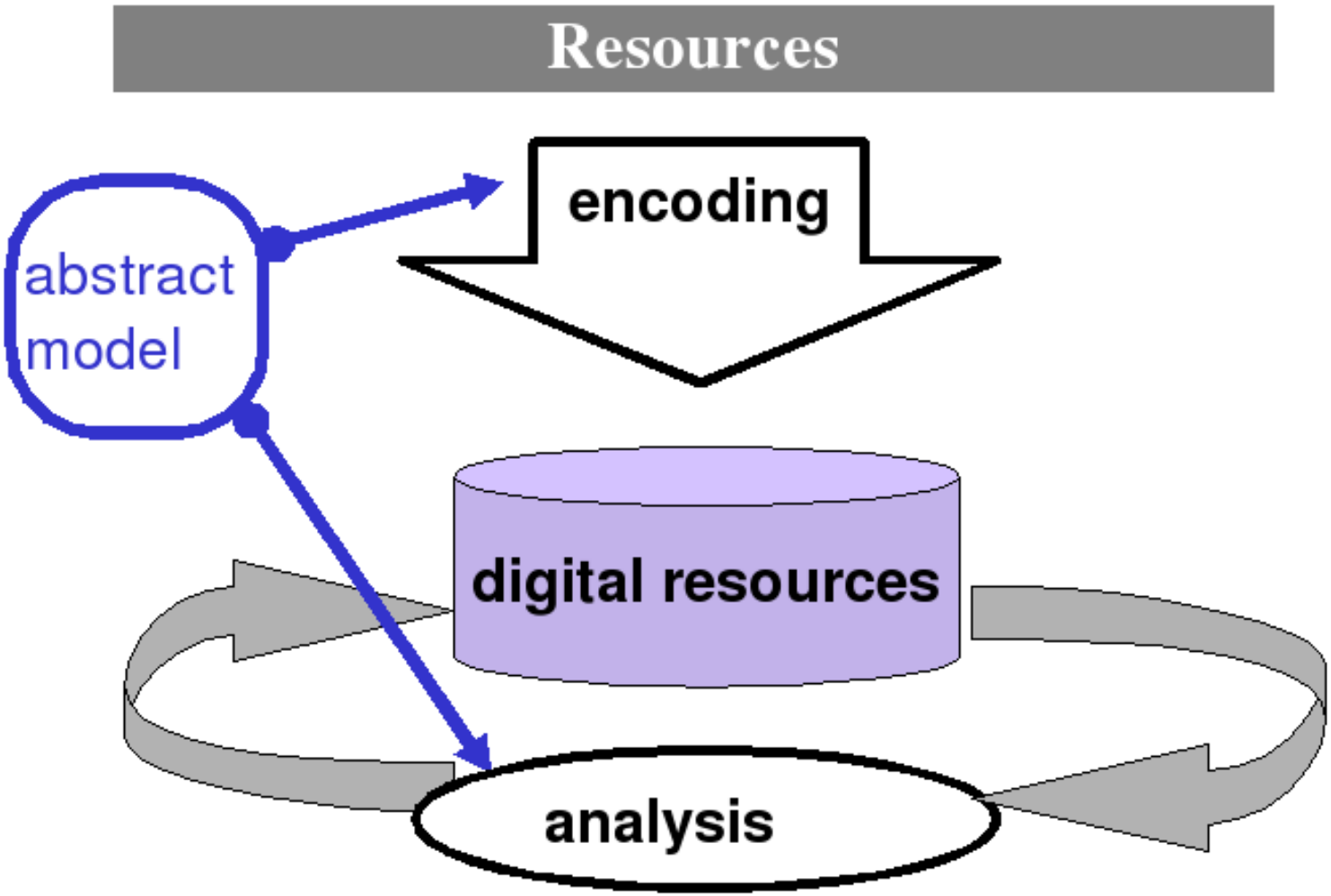
- because we need to interchange resources
  - between people
  - (increasingly) between machines
- because we need to integrate resources
  - of different media types
  - from different technical contexts
- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data

Relevance of the TEI

Why would you want those things?

- because we need to interchange resources
  - between people
  - (increasingly) between machines
- because we need to integrate resources
  - of different media types
  - from different technical contexts
- because we need to preserve resources
  - cryogenics is not the answer!
  - we need to preserve metadata as well as data

The virtuous circle of encoding



The scope of intelligent markup

Even within the original scope of the TEI we have

- basic structural and functional components
- diplomatic transcription, images, annotation
- links, correspondence, alignment
- data-like objects such as dates, times, places, persons, events (*named entity recognition*)
- meta-textual annotations (correction, deletion, etc)
- linguistic analysis at all levels
- contextual metadata of all kinds
- ... and so on and so forth

Is it possible to delimit encyclopaedically all possible kinds of markup?

Reasons for attempting to define a common framework

- re-usability and repurposing of resources
- modular software development
- lower training costs
- 'frequently answered questions' — common technical solutions for different application areas

The TEI was *designed* to support multiple views of the same resource

Being a good digital citizen

- XML implies Unicode; but the TEI also provides markup for non-Unicode characters and glyphs
- TEI schemas can be generated for
  - Traditional XML DTD language
  - ISO RELAX NG language
  - W3C Schema Language
- TEI content models use (an interoperable subset of) RELAX NG syntax
- TEI datatypes are defined in terms of W3C datatypes
- All linking and pointing uses W3C standards
- Additional constraints may be expressed in ISO Schematron or similar
- Hooks are provided for mapping to other ontological frameworks
- Namespaces are fully supported

For example

Embedding SVG within TEI:

```
<figure>
 <svg xmlns="http://www.w3.org/2000/svg"
  width="6cm" height="5cm" viewBox="6 3 6 5">
  <ellipse xmlns="http://www.w3.org/2000/svg"
   style="fill:
#ffffff" cx="9.75" cy="6.35" rx="2.75" ry="2.35"/>
 </svg>
</figure>
```

A user-defined attribute:

```
<div
   xmlns:my=http://www.example.org/ns/nonTEI>
 <p n="12" topic="rabbits">Flopsy, Mopsy, Cottontail, and
Peter...</p>
</div>
```

James Clark's *onvdl* processer validates against multiple namespace schemas

Conformance issues

A document is *TEI Conformant* if and only if it:

- is a well-formed XML document
- can be validated against a *TEI Schema*, that is, a schema derived from the TEI Guidelines
- conforms to the TEI Abstract Model
- uses the *TEI Namespace* (and other namespaces where relevant) correctly
- is documented by means of a TEI Conformant *ODD file* which refers to the TEI Guidelines

or if it can be transformed automatically using some TEI-defined procedures into such a document (it is then considered *TEI-conformable*).

Standardization should not mean 'Do what I do', but rather 'Explain what you do in terms I can understand'

TEI Default Text Structure

All TEI documents are structured in a particular manner. This section attempts to describe the different variations on this as briefly as possible.

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)

```
<teiCorpus>
  <teiHeader>
<!-- required -->
  </teiHeader>
  <TEI>
<!-- required -->
  </TEI>
</teiCorpus>
```

TEI basic structures (2)

```
<TEI>
 <teiHeader>
<!-- required -->
 </teiHeader>
 <facsimile>
<!-- optional, new in TEI P5 -->
 </facsimile>
 <text>
<!-- required if no facsimile -->
 </text>
</TEI>
```

<text>

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
- a unitary text contains
  - optional front matter
  - <body> (required)
  - optional back matter

<text>

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
- a unitary text contains
  - optional front matter
  - <body> (required)
  - optional back matter

Composite texts

A composite text contains

- optional front matter
- <group> (required)
- optional back matter

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

<group> tags may self-nest.

TEI text structure (1)

```
<text>
 <front>
<!-- optional -->
 </front>
 <body>
<!-- required -->
 </body>
 <back>
<!-- optional -->
 </back>
</text>
```

TEI text structure (2)

```
<text>
 <front>
<!-- ... -->
 </front>
 <group>
  <text>
   <body>
    <p>...</p>
   </body>
  </text>
 </group>
 <back>
<!-- ... -->
 </back>
</text>
```

Another Grouped Text Example

```
<TEI>
 <teiHeader>
<!-- header information for the whole collection -->
 </teiHeader>
 <text>
<!-- optional front matter -->
   <group>
    <text>
<!-- optional front matter -->
     <body>
<!-- First Body -->
     </body>
    </text>
    <text>
<!-- optional front matter -->
     <body>
<!-- Second Body-->
     </body>
    </text>
   </group>
 </text>
</TEI>
```