ENRICH training slides

Lou Burnard, James Cummings, Arianna Ciula, Matthew Driscoll, Sebastian Rahtz

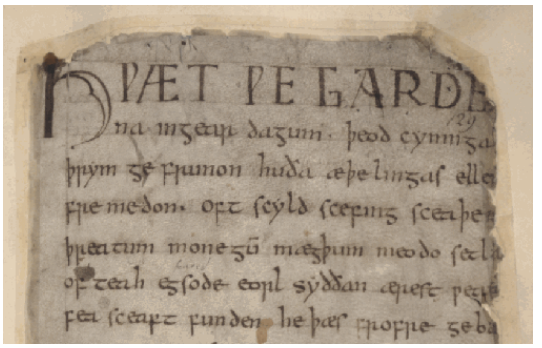
September 22, 2008

What is XML markup for?

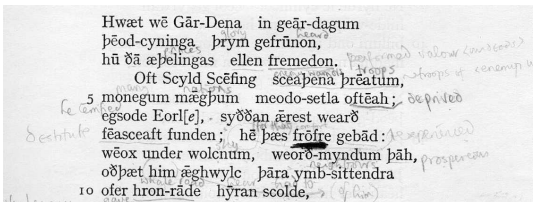
TEI @ Oxford

September 2008

What's in a text ?



Is this the same text?



Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceaþena þrēatum,
monegum mægþum meodo-setla oftēah;
egsode Eorl[e], syððan ærest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
ofer hron-rāde hȳran scolde,

The ontology of text

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in its author's intentions?

A "text" is an abstraction, created by or for a community of readers. Markup encodes and makes concrete such abstractions.

Encoding of texts

- Texts are more than sequences of encoded glyphs
  - They have structure and content
  - They also have multiple readings
- Encoding, or markup, is a way of making these things explicit

Only that which is explicit can be reliably processed

What's the point of markup?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

It's (usually) more useful to markup what we think things *are* than what they *look like*

Markup as a scholarly activity

- The application of markup to a document can be an intellectual activity
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor
- There is (almost) no such thing as neutral markup -- all of it involves interpretation
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself
- Good textual encoding is never as easy or quick as people would believe
- Detailed document analysis is needed before encoding for the resulting markup to be useful

What does markup capture?

Compare

```
<lb/>
<hi rend="dropcap">H</hi>
<g ref="#WYNN"/>ÆT WE GARDE
<lb/>na in gear-dagum þeod-cyninga
<lb/>þrym gefrunon, hu ða æþelingas
<lb/>ellen fremedon. oft scyld scefing sceaþe
<supplied>na</supplied>
<lb/>þreatum, moneg<ex>um</ex> mægþum meodo-setl
<supplied>a</supplied>
<lb/>of<damage>
  <desc>blotted</desc>
</damage>teah ...
```

and

```
<lg>
  <l>Hwæt! we Gar-dena in gear-dagum</l>
  <l>þeod-cyninga þrym gefrunon,</l>
  <l>hu ða æþelingas ellen fremedon,</l>
</lg>
<lg>
  <l>Oft Scyld Scefing sceaþena þreatum,</l>
  <l>monegum mægþum meodo-setla ofteah;</l>
  <l>egsode Eorle, syððan ærest wearþ</l>
  <l>feasceaft funden...</l>
</lg>
```

Some alphabet soup

SGML	Standard Generalized Markup Language
HTML	Hypertext Markup Language
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
DTD	Document Type Definition (or Declaration)
CSS	Cascading Style Sheet
Xpath	XML Path Language
XSLT	eXtensible Stylesheet Language - Transformations
XQuery	XML Querying
RELAXNG	Regular Expression Language for XML (New Generation)

Oh, and then there's also TEI, the *Text Encoding Initiative*

XML: what it is and why you should care

- XML is structured data represented as strings of text
- XML looks like HTML, except that:-
  - XML is extensible
  - XML must be well-formed
  - XML can be validated
- XML is application-, platform-, and vendor- independent
- XML empowers the content provider and facilitates data integration

XML terminology

An XML document may contain:-

- elements, possibly bearing attributes
- processing instructions
- comments
- entity references
- marked sections (CDATA, IGNORE, INCLUDE)

An XML document must be well-formed and may be valid

The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- Attribute names are predefined for a given element; values can also be constrained

Representing an XML tree

- An XML document is encoded as a linear string of characters
- It begins with a special processing instruction
- Element occurrences are marked by start- and end-tags
- The characters < and & are Magic and must always be "escaped" if you want to use them as themselves
- Comments are delimited by <!- - and - ->
- CDATA sections are delimited by <![CDATA[and]]>
- Attribute name/value pairs are supplied on the start-tag and may be given in any order
- Entity references are delimited by & and ;

A complete XML document

```
<?xml version="1.0"?>
<greetings xmlns="http://www.example.com/ns">
<hello type="fulsome">hello world!</hello>
</greetings>
```

- The XML declaration
- Namespace declaration
- The root element of the document itself
- Other elements and content
- Attribute and value

The XML declaration

```
<?xml version="1.0" encoding="iso-8859-1"?>
```

An XML document must begin with an XML declaration which does two things:

- specifies that this *is* an XML document, and which version of the XML standard it follows
- may specify a different character encoding for the document — if the default, and recommended, encoding UTF-8 is not being used

Namespace declarations

An XML document may include elements declared in different name spaces.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:math="http://www.mathml.org">
```

- a namespace declaration associates a namespace prefix with an external URI-like identifier
- the default namespace *may* be declared using a `xmlns`
- other name spaces must all use a specially declared prefix
- All TEI documents are declared within the TEI namespace
- The xml namespace is available in all XML documents; TEI uses it for global attributes *@xml:id* and *@xml:lang*

The Doctype Declaration

You may sometimes find an optional "Document Type" declaration:

```
<?xml version="1.0" ?>
<!DOCTYPE greeting SYSTEM "greeting.dtd []">
```

- The DTD is one way of associating the document with its schema (but is not used by W3C or RELAXNG for this purpose)
- The DTD subset is used to provide declarations additional to those in the schema, for example for external files
- The DTD subset may be internal, external, or both

DTDs are now considered old-fashioned — RELAXNG or W3C schemas are preferred.

XML syntax: the small print

What does it mean to be well-formed?

1. there is a single root node containing the whole of an XML document
2. each subtree is properly nested within the root node
3. names are always case sensitive
4. start-tags and end-tags are always mandatory (except that a combined start-and-end tag may be used for empty nodes)
5. attribute values are always quoted

Note: You can be valid in addition to being well-formed. This means you obey the rules of a specified schema, such as the TEI.

Test your XML knowledge

- Which are correct?
  - `<seg>some text</seg>`
  - `<seg><foo>some</foo> <bar>text</bar></seg>`
  - `<seg><foo>some <bar></foo> text</bar></seg>`
  - `<seg type="text">some text</seg>`
  - `<seg type='text'>some text</seg>`
  - `<seg type=text>some text</seg>`
  - `<seg type = "text">some text</seg>`
  - `<seg type="text">some text<seg/>`
  - `<seg type="text">some text<gap/></seg>`
  - `<seg type="text">some text< /seg>`
  - `<seg type="text">some text</Seg>`

Live long and prosper! Lessons from the TEI

TEI @ Oxford

September 2008

1986 was a long time ago…

- The first computer virus – Brain – appears, in the USA
- Construction of the channel tunnel begins
- The Soviet Union launches space station Mir
- Disaster at Chernobyl
- Olaf Palme assassinated
- Records of the year: Raising Hell (Run DMC)… Graceland (Paul Simon)… Группа крови (Виктор Цой)

…but we used computers then

- Corpus linguistics
- Databases on CD ROM
- Largescale lexical resources already existed (eg TLF, TLG, LASLA…)
- Digital lexicography (e.g. OED)
- Document management systems (e.g. TeX, Scribe, tRoff..)
  - some proprietary (and expensive), some research
- Text archives
- Hypertext theory

But there was no world wide web and not many desktop pcs…

Birth 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"

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 Excel
- 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?

Is the TEI still relevant?

- With XML everyone can create their own markup system and still share data!
- In the Semantic Web, XML systems will all understand each other's data!

If we have

- historical data marked up with a Historical Markup Language
- linguistic data marked up with a Linguistic Markup Language
- metadata marked up with a Metadata Markup Language

how will we integrate resources or ask interesting questions?

Haven't we been here before?

Relevance 1

The TEI provides

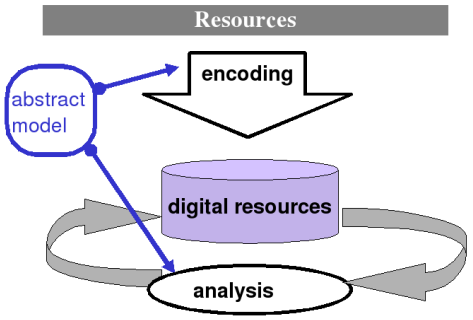
- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) resources…
- … which can be enriched and personalized in highly idiosyncratic or specialised ways
- a very rich library of existing specialised components
- an integrated suite of standard stylesheets for delivering schemas and documentation in various languages and formats
- a large and active open source style user community

Relevance 2

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

Old Skool TEI

- A traditional (if large) research project with soft funding, driven by academic curiosity
- a codification of best practice, with no formal maintenance method
- uncertain licencing and development practices
- perceived as unmanageably complex except by the priesthood — or simultaneously as too simple for real scholarly work
- lack of specific tools to *do* something with a TEI text
- failure to market the advantages of rich markup

TEI New

- Proper open source licence, with visible development on Sourceforge
- Architecture rethought to facilitate expansion and integration with other systems
- Self documenting, each release fully validated, delivered using standard mechanisms
- Publicly available processing tools managed together with the Guidelines
- Active developer community, wiki, etc. Test files, exemplars, regular updates…
- New governance structure, new tools, new modules…

Three important things about TEI P5

1. Being a good digital citizen:
   - Support for multiple schema languages and namespaces
   - Reliance on XML, and hence on Unicode
   - Validation of attributes and datatyping
   - Use of W3C pointers and paths
2. Making it flexible:
   - ODD: a single specification language for developers, users, and teachers, integrating schema and documentation;
   - Verifiable conformance
3. Old annoyances removed and some new topics added

One Specification Language

- A set of TEI documents is described by an ODD, which is itself a TEI document that combines:
  - references to existing declarations
  - formal declarations for elements and attributes
  - documentation and usage notes
- Underlying this:
  - a conceptual model which abstracts from specific elements to generic classes
  - a modular architecture for combining sets of definitions
- specifications are chainable; modifications are written in ODD with ODD as input and output
- Roma is one interface to this: there will be others

For example

An ODD file is a valid TEI document, containing descriptive prose, and a `<schemaSpec>` element to define the schema it documents

```
<div>
 <head>Our Project Manual</head>
 <p>In this project we use the basic TEI structures
   with a few minor modifications to exclude
   elements we do not need</p>
 <schemaSpec ident="TEI-minimal" start="TEI">
  <moduleRef key="tei"/>
  <moduleRef key="header"/>
  <moduleRef key="core"/>
  <moduleRef key="textstructure"/>
<!-- We don't need these drama elements: -->
  <elementSpec ident="sp" mode="delete" module="core"/>
  <elementSpec ident="speaker" mode="delete" mod-
ule="core"/>
 </schemaSpec>
</div>
```

Support for many schema languages

- TEI schemas can be generated for
  - XML DTD language
  - ISO RELAX NG language
  - W3C Schema Language
- Content models are defined using RELAX NG syntax
- Datatypes are defined in terms of W3C datatypes
- Some facilities (e.g. alternation, namespaces) cannot be expressed in DTD
- Additional constraints can be expressed in Schematron

Two reasons why standards fail

- The theory is not yet ripe
- The "not invented here" attitude: the community of users is too diverse

Coping with partially-baked ideas

In a TEI ODD, you can …

- constrain the domain of a value list
- enforce Schematron rules about e.g. co-dependency
- provide new elements in your own namespace
- remove (non-mandatory) child elements

New elements

A schema is a grammar. How can you add new terminals to an existing syntax?

- All content models are expressed indirectly, by reference to element classes rather than elements
- Hence adding a new element is simply a matter of saying which class/es it belongs to

The TEI schema is also enriched with semantics. How can you explain what a new element means?

- Class membership also conveys some semantics
- ODD includes detailed documentation

Coping with the NIH Syndrome

- TEI P5 has extensive I18N features for translation of …
  - schema objects
  - schema documentation
- See *Roma* at `http://www.tei-c.org/Roma/`
- TEI is hospitable to other namespaces
  - You can use SVG for graphics, MathML for math, Word Table markup if you like
  - (but note this doesn't solve the Other Overlap Problem)
- ODD also includes an `<equiv>` element for mapping to external ontologies

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" my:topic="rabbits">Flopsy, Mopsy, Cottontail,
and Peter…</p>
</div>
```

An NVDL processor can validate a document using 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 TEI-conformable)

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

Evolution works!

1. Make modifications in your own namespace
2. Document them in an ODD
3. Propose them to the TEI Council as amendments or feature requests
4. TEI P5 now has a 6 month release cycle…

Visit `http://www.tei-c.org` for more background info

Visit `http://tei.sf.net` to download

Using the basic TEI structural elements

TEI @ Oxford

September 2008

TEI Infrastructure

- The TEI encoding scheme consists of a number of modules
- These declare XML elements and their attributes
- An element's declaration assigns it to one (or more) model classes
- Another part declares its possible content and attributes with reference to these classes
- This indirection allows strength and flexibility
- It makes it easy to add/exclude new elements by referencing existing classes

What is a module?

- A convenient way of grouping together a number of element declarations
- These are usually on a related topic or specific application
- Most chapters focus on elements drawn from a single module, which that chapter then defines
- A TEI Schema is created by selecting modules and add/removing elements from them as needed

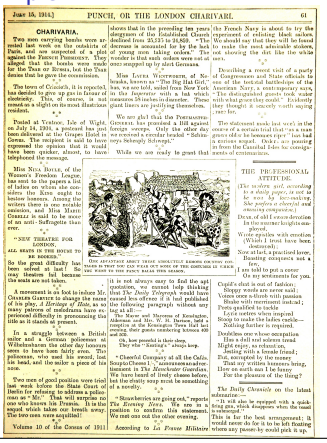
Modules

Module name	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 Elements
tei	The TEI Infrastructure
textcrit	Critical Apparatus
textstructure	Default Text Structure
transcr	Representation of Primary Sources
verse	Verse

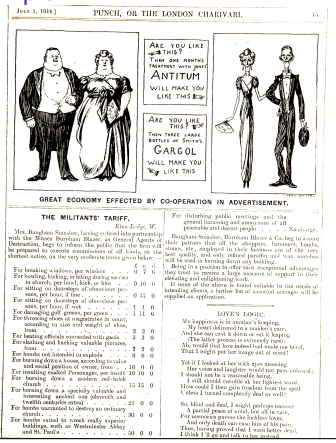
The Imaginary Punch Project

- Punch is a famous English humorous journal, published regularly between 1841 and 1992: see `http://www.punch.co.uk/historyofpunch.html`.
- A project plans to make available fully marked up texts of the journal, in conjunction with page images...
  - for social historians
  - for librarians
  - for linguists
- How will the TEI help? And which parts of the TEI will we use?

Example 1



Example 2



Example 3

Looking at Punch, what do we need to mark up?

- issue information and page number for reference purposes
- "chunks" or divisions of text, which may contain a picture, a poem, some prose, some drama, or a combination
- within the chunks, we can identify formal units such as
  - a picture, a caption
  - stanzas, lines
  - paragraphs
  - speeches and stage-directions
- and more...

TEI tags for the high level structure

We will treat each issue as a single `<text>` element, and each identifiable chunk within it as a `<div>` element of a particular type (e.g. cartoon, verse, prose)
For example, page 1 has two divisions,

```
<pb n="1"/>
<div type="cartoon">....</div>
<div type="poem">
  <head>Progress</head>....
</div>
```

page 2 also has two, of different types:

```
<pb n="2"/>
<div type="prose">
  <head>The enchanted castle</head>....
</div>
<div type="snippet">
  <head>Correspondence</head>....
</div>
```

Why divisions rather than pages?

Because a division can start on one page (page 5 for example) and finish on another (page 6)
We use an empty element`<pb>` to mark the boundary between pages, rather than enclosing each page in a `<div type="page">`.

```
<pb n="5"/>
<div type="cartoon">...</div>
<div type="review">
  <head>Egypt in Venice</head>...
<pb n="6"/>
...
</div>
<div type="cartoon">...</div>
<div type="verse">
  <head>Enigma</head>...
</div>
<div type="snippets">...</div>
```

The sequence in which divisions appear is rather arbitrary.

Divisions can contain divisions...

```
<div type="snippets">
  <div type="snippet">Curiously....Chancellor</div>
  <div type="snippet">Men for the Antarctic...  Canadians</div>
</div>
```

- TEI also provides division elements with names that indicate their degree of nesting (`<div1>`, `<div2>` etc.) which some people prefer
- Divisions must always tessellate: once "down" a level, you cannot pop "up" again within the same division.

Floating text

As mentioned above, s must tesselate over the entire text

```
<div1>
  <p> ... </p>
  <div2>
    <p> ... </p>
  </div2>
  <div2>
    <p> ... </p>
  </div2>
</div1>
```

is valid but

```
<div1>
  <p> ... </p>
  <div2>
    <p> ... </p>
  </div2>
  <p> ... </p>
</div1>
```

is *not valid*.

A special <floatingText> element is available for "interruptions"

What are divisions made of?

(apart from other smaller divisions)

- <head> (heading)
- <p> (paragraph)
- <sp> (speech, contains any of the foregoing, also <stage> and <speaker>)
- <list> (contains <head>, <label>, <item>)
- <table>, (contains <row> containing <cell>) ...
- <l> (verse line) optionally grouped into <lg> (line group) stanzas
- <figure> (contains <graphic>, <figDesc>, <head>...)

For example....

Page 3 contains a figure and a dialogue...

```
<div type="cartoon">
  <figure>
    <head>When the ships come home</head>
    <figDesc>A man in Turkish dress lounges on a sofa,
        smoking a cigarette and consulting a book
        labelled "Naval ledger". Another man, in
        traditional Greek costume, stands beside him,
        also reading a notebook.</figDesc>
    <graphic url="Punch/XML/Graphics/003.png"/>
  </figure>
  <sp>
    <speaker>Greece.</speaker>
    <p> Isn't it time we started fighting again?</p>
  </sp>
  <sp>
    <speaker>Turkey.</speaker>
    <p> Yes, I daresay. How soon could you begin?</p>
  </sp>
  <sp>
    <speaker>Greece.</speaker>
    <p> Oh, in a few weeks.</p>
  </sp>
  <sp>
    <speaker>Turkey.</speaker>
    <p> No good for me. Shan't be ready till
        the autumn.</p>
  </sp>
</div>
```

For example...

The militants' tariff (on Page 15) contains headings, paragraphs, and a table...

```
<div type="prose">
  <head>THE MILITANTS' TARIFF.</head>
  <head rend="right">Etna Lodge, W.</head>
  <p>Mrs. Bangham Smasher, having entered into partnership with the
    Misses Burnham Blazer, as General Agents of Destruction, begs to
    inform the public that the firm will be prepared to execute
    commissions of all kinds, at the shortest notice, on the very
    moderate terms given below: —
  </p>
  <table>
    <row role="label">
      <cell/>
      <cell>£</cell>
      <cell>s.</cell>
      <cell>d.</cell>
    </row>
    <row>
      <cell>For breaking windows, per window ...</cell>
      <cell>0</cell>
      <cell>7</cell>
      <cell>6</cell>
    </row>
    <row>
      <cell>For howling, kicking, or biting during service
          in church, per howl, kick, or bite ...</cell>
      <cell>0</cell>
      <cell>10</cell>
      <cell>6</cell>
    </row>
    <row>
```

Global attributes

Some features (potentially) apply to everything:

- identity
- language
- rendition

TEI provides global attributes for these:

- *@xml:id* provides a unique identifier for any element;
- *@n* provides a name or number for any element
- *@xml:lang* specifies the language of any element, using an ISO standard code
- *@rend* and *@rendition* provide ways of specifying the visual appearance (rendition) of any element

For example...

Egypt in Venice (on Page 05) begins with two headings, one in French....

```
<div type="prose" xml:lang="en" xml:id="I1914-07-01_05_02">
  <head>Egypt in Venice.</head>
  <head xml:lang="fr" rend="it">"La Légende de Joseph."</head>
  <p>Those who know the kind of attractions that the
    Russian ballet offers in so many of its themes ....</p>
</div>
```

Each stanza of the poem on page 10 has a last line which is significantly indented:

```
<lg>
  <l>There were eight pretty walkers who went up a hill;</l>
  <l>They were Jessamine, Joseph and Japhet and Jill,</l>
  <l>And Allie and Sally and Tumbledown Bill,</l>
  <l rend="indent">And Farnaby Fullerton Rigby.</l>
</lg>
```

Macrostructure 1

All the issues of *Punch* for one year make up a volume. We could regard the volume as a single <text>, and each issue as a within it. Or we could use the <group> element:

```
<text xml:id="v147">
  <front>
<!-- introductory materials for volume 147 here -->
  </front>
  <group>
    <text xml:id="I1914-07-01">
      <body>
<!-- first issue (1 July) -->
      </body>
    </text>
    <text xml:id="I1914-07-15">
      <body>
<!-- second issue (15 July) -->
      </body>
    </text>
<!-- etc... -->
  </group>
  <back>
<!-- volume index, appendix etc. -->
  </back>
</text>
```

Macrostructure 2

As well as the texts, we have detailed metadata about each volume, and images of its pages. These are the three parts of a canonical TEI document:

```
<TEI>
  <teiHeader>
<!-- required; provides metadata -->
  </teiHeader>
  <facsimile>
<!-- the text, represented in image form -->
  </facsimile>
  <text>
<!-- the text, transcribed and marked up -->
  </text>
</TEI>
```

Macrostructure 3

If many such documents are grouped together to form a corpus (rather than a collection), it may be useful to factor out the metadata they have in common:

```
<teiCorpus>
  <teiHeader>
<!-- shared metadata -->
  </teiHeader>
  <TEI>
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
  <TEI>
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```

What kinds of metadata?

For the *Punch Project* and for any other comparable project, we will need a place for such information as

- identification of the resource itself ("what is this thing?")
- statements of responsibility ("who did what when?")
- indication of source ("what was this derived from?")
- publication statement ("how is this item distributed and by whom?")
- declaration of encoding practice ("what do the codes we added mean?")

The TEI Header supports all these, and more.

The TEI Header

The TEI header was designed with two goals in mind

- needs of bibliographers and librarians trying to document 'electronic books'
- needs of text analysts trying to document 'coding practices' within digital resources

On the one hand, the Librarian's header

- uses standard bibliographic concepts
- respects established mappings to other such records (e.g. MARC)
- has a preference for structured data over loose prose

On the other, Everyman's header

- Supports a (potentially) huge range of very miscellaneous information, organized in fairly ad hoc ways -
- Unpredictable combinations of narrowly encoded documentation systems and loose prose descriptions

TEI Header Structure

The TEI header has four main components:

- <fileDesc> (file description) contains a full bibliographic description of an electronic file.
- <encodingDesc> (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
- <revisionDesc> (revision description) summarizes the revision history for a file.
- <profileDesc> (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting. (just about everything not covered in the other header elements

Only <fileDesc> is required; the others are optional.

Simple TEI Header for Punch Project

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Punch, or the London Charivari, Vol. 147, July 1, 1914</title>
    </titleStmt>
    <publicationStmt>
      <idno type="gutenberg">24357</idno>
      <availability>
        <p>This text is freely available for re-use
           under US and UK law, consult your local
           legal restrictions if elsewhere.</p>
      </availability>
    </publicationStmt>
    <sourceDesc>
      <p>This text is a TEI version of a Project Gutenberg
         text originally located at <ptr
        target="http://www.gutenberg.org/dirs/2/4/3/5/24357/"/>.
         As per their license agreement we have removed all
         references to the PG trademark.</p>
    </sourceDesc>
  </fileDesc>
  <revisionDesc>
    <change when="2008-07-26T23:49:55.968+01:00"/>
  </revisionDesc>
</teiHeader>
```

Below the paragraph...

Within the elements already introduced, TEI offers plenty of scope for mark-up of smaller components. For example:

- boundaries, such as page, column, or line breaks
- highlighting, emphasis and quotation
- editorial changes such as correction, normalization etc.
- names, numbers, dates, addresses...
- links and cross-references
- notes, annotation, indexing
- graphics
- bibliographic citations
- words and other analyses

Highlighting

By highlighting we mean any combination of typographic features (font, size, hue, etc.) which distinguishes the highlighted text from its surroundings. This may be for many reasons...

- to mark foreign, archaic, technical usages
- for emphasis when spoken
- to show something is not part of the text.. (e.g. cross references, titles, headings)
- or is attributed to some other agency inside or outside the text (e.g. direct speech, quotation)

TEI provides both a generic <hi> tag and a large number of specific ones...

A few highlighting examples

- <hi> (highlighted: reason unknown or unimportant)

```
<p>[The rest of this communication is
omitted owing to considerations of
space.—<hi rend="sc">Ed</hi>.]</p>
```

- <emph> (emphasized)

```
<said>'E won't bite yer <emph>if you buy 'im</emph> guv'ner.</said>
```

- <title> and <foreign>:

```
<p>
  <foreign xml:lang="fr">À propos</foreign> of Oxford, it is a
question whether that extremely amusing book
<title>Verdant Green</title> is still much read by freshers.
</p>
```

- <distinct> (linguistically marked)

```
But then I remind myself
that the Russian ballet is nothing if not
<distinct>bizarre</distinct>
```

Quotation

Quotation marks can similarly be used to set off text for many reasons:

- <q> (used if the reason is unknown or unimportant)
- <said> (speech or thought)
- <quote> (attributed to an external source)
- <mentioned> and <soCalled> (nuances of narrative status)

```
<p>
  <said who="#Celia">I know a lovely tin of potted
    grouse,</said> said Celia, and she went off to cut some sandwiches.
</p>
```

```
<head>How to utilise the art of <soCalled>suggestion</soCalled>
</head>
<head>The Doctor, six down at the turn,
<soCalled>suggests</soCalled> to his opponent that
they are playing croquet, and wins by two and one.</head>
```

Quotation (continued)

Note that these elements can nest within one another:

```
<p>The poet returned to his work. <said>
   <quote>In
        tooth and claw,</quote>
  </said> he muttered to himself,
<said>
   <quote>In tooth and claw.</quote>
  </said>
</p>
```

Editorial intervention

As a simple example, consider: 'Excuse me sir, but would you like to buy a nice little dawg?' on page 6.
We can:

- use <orig> to show that "dawg" is what it says, even though this is a nonstandard spelling
- use <reg> to show that "dog" is an editorially-supplied regularisation of what it says
- or provide both within a <choice> element to say either is a valid encoding:

```
...a nice little
<choice>
  <orig>dawg</orig>
  <reg>dog</reg>
</choice>?
```

Names of persons, places, things...

- <name> (a name in the text, contains a proper noun or noun phrase)
- <rs> (a general-purpose name or referencing string)
- <title> (any form of title)

The *@type* attribute is useful for categorizing these, and they both also have *@key*, *@ref*, and *@nymRef* attributes.

Examples of names

Using *@type* to distinguish personal from geographic names:

```
<p>The scene opens at a party given by
<name type="person">Potiphar</name> in
<name type="place">Venice</name>.  </p>
```

Using *@key* and *@ref* to de-reference names:

```
<p>
  <label>Business done. </label>—The Commons
still harping on the Budget.
<name
     type="person"
     ref="http://en.wikipedia.org/wiki/Timothy_Michael_Healy">
     Tim Healy</name> enlivened proceedings by vigorous personal attack
on <q>the most reckless and incapable
  <rs key="LLG">Chancellor of the Exchequer</rs>
     that ever sat on the Treasury Bench. </q>
  <name key="LLG">Lloyd George's</name>
retort courteous looked forward to with interest.
</p>
```

Dates

- <date> contains a date and time in any format
- For processing it is convenient to add a normalized version, using the *@when* attribute
- Uncertain dates and times, and ranges, can be indicated by other attributes: *@notBefore*, *@notAfter*, *@from@to*

```
<p>House of Commons, <date when="1914-06-22">Monday, June 22, 1914</date>. </p>
<p>
  <date notAfter="1914-06-01" notBefore="1914-03-01">Sunday, a month
ago,</date> was hot.
</p>
```

Cross references

A cross reference is a link from one point in a text (the source) to another (the target).
TEI provides generic elements <ptr> and <ref> for this purpose.
If the linking text can be automatically generated use <ptr>; otherwise use <ref>.
The source is the location of the <ptr> or <ref>; the target is specified by the *@target* attribute, in the form of a URI reference.

```
See <ref target="#Section12">section 12 on page 34</ref>.
```

```
See <ptr target="#Section12"/>.
```

Bibliographic Citations

TEI provides special elements for bibliographic citations or references:

- <bibl> (loosely structured)
- <biblStruct> (standard bibliographic structure)
- <listBibl> (encloses a bibliography)

These are typically used in preparing bibliographies, or in footnotes. But even in Punch, there are examples.

Simple <bibl> Example

In Punch, bibliographic citations are usually associated with a a quotation from another paper:
The<cit> element groups the two:

```
<cit>
  <quote>It was the time when Henry III. was
     batting with Simon de Montfort and his
     Barons.</quote>
  <bibl>
    <title>Straits Times.</title>
  </bibl>
</cit>
```

Embedded notes

Notes, whether appearing in the original source, or added by an editor, can be marked using the <note> element.
We might use this to add biographical details to the Punch transcriptions:

```
<p>By-the-by, it is denied that
Sir <name rend="sc">Joseph Beecham</name>
  <note>Sir Joseph Beecham, 1st Baronet
     (8 June 1848 - 23 October 1916)...</note>.
was in any way responsible for the Government's
"Pills for Earthquakes," by which it was hoped to
avert the Irish crisis.</p>
```

<note> has attributes *@place* and *@resp*

Linked notes

Since we have several references to the same person, it might be better to put the notes elsewhere and point to them from the names:

```
<div type="notes">
  <note xml:id="BEECHJO">Sir Joseph Beecham, 1st Baronet (8 June 1848 -
     23 October 1916) the eldest son of Thomas Beecham (1820-1907) played a
     large part in the growth and expansion of his father's medicinal pill
     business which he joined in 1866....</note>
<!-- other notes -->
</div>
<div type="snippets">
  <p>...  Both Earl <name rend="sc">Beauchamp</name>
     and <name>Sir <ref target="#BEECHJO">Joseph Beecham</ref>
    </name> appear
     in the recent Honours List.</p>
  <p>By-the-by, it is denied that Sir <name rend="sc" ref="#BEECHJO">Joseph
       Beecham</name> was in any way responsible...</p>
</div>
```

Could also use specialised <person> element, in this case.
"Elsewhere" can be *anywhere* on the Internet...

Names, People, and Places

TEI @ Oxford

September 2008

What's in a name?

- We've already met <name> and <rs> for any form of name or referring string.
- The namesdates module also provides specialisations of these: <persName>, <placeName>, and <orgName>
- Each can be further decomposed
- They can also be associated with a named entity
- (Names are also entities)

Personal Names

For example...

- <persName> (personal name) a noun referring to a person ... equivalent to <name type="person">
- <surname> a family (inherited) name
- <forename> a forename, given or baptismal name
- <roleName> a name component indicating a particular role or position in society
- <addName> (additional name) nickname, epithet, alias, or any other descriptive phrase used within a personal name
- <nameLink> a connecting phrase or link used within a name but not regarded as part of it

```
<persName>
  <forename type="first">Inês</forename>
  <forename type="matronymic">Barroca</forename>
  <surname>Rahtz</surname>
</persName>
```

Names as referents (1)

In a text we might find the same person referred to on different occasions in any number of different ways:

```
...<persName>Clara Schumann</persName>....
<persName>Clara</persName>
....
<persName>Frau Schumann</persName>
```

All of these names refer to the same entity
We can use an attribute on any naming element to specify which entity is being referenced:

- *@key* if we are supplying an externally-defined code for the entity
- *@ref* if we are pointing to a definition of the entity

Names as referents (2)

For example:-

```
...<persName ref="#CS">Clara Schumann</persName>....
<persName ref="#CS">Clara</persName>
....
<persName key="CS123">Frau Schumann</persName>
<!-- ...  elsewhere -->
<person xml:id="CS" sex="2">
  <persName xml:lang="de">
    <forename type="first">Clara</forename>
    <forename type="middle">Josephine</forename>
    <surname type="maiden">Wieck</surname>
    <surname type="married">Schumann</surname>
  </persName>
</person>
```

The thing itself (1)

TEI provides special-purpose elements for maintaining structured information about named entities (as well as their names):

- <person>, <place>, <event>
- may be grouped into <listPerson>, <listPlace>, (and soon <listEvent>)
- relationships can also be modelled, explicitly using <relation> or implicitly by context

```
<person xml:id="VM1893" sex="1">
  <persName xml:lang="ru">Владимир Владимирович Маяковский</persName>
  <persName xml:lang="fr">Wladimir Maïakowski</persName>
  <birth when="1893-07-19">7 July (OS) 1893,
<placeName ref="#BGDT" xml:lang="en">Baghdati, Georgia</placeName>
  </birth>
  <death when="1930-04-14"/>
  <occupation>Poet and playwright, among the foremost representatives of
    early-20th century Russian Futurism.</occupation>
<!-- ... -->
</person>
```

Traits, states, and events

The scope of elements one might record for a named entity is *large*. The TEI provides three generic elements, and some specific ones. We identify three main classes of information:

- characteristics or traits which do not, by and large, change over time
- characteristics or states which hold true only at a specific time
- events or incidents which may lead to a change of state or, less frequently, trait

For a person, typical traits are such things as <faith>, <sex>, <socEcStatus>; typical states are such things as <occupation>, <residence>, <education>; typical events are such things as <birth> and <death>.

Personal Relationships

- <relationGrp> (relation group) provides information about relationships identified amongst people, places, and organizations
- <relation> (relationship) describes any kind of relationship or linkage amongst a specified group of participants
  - *@name* supplies a name for the kind of relationship of which this is an instance
  - *@active* identifies the 'active' participants in a non-mutual relationship, or all the participants in a mutual one
  - *@mutual* supplies a list of participants amongst all of whom the relationship holds equally
  - *@passive* identifies the 'passive' participants in a non-mutual relationship

Example

```
<person xml:id="jsbach" sex="1">
 <persName>Johann Sebastian Bach</persName>
</person>
<person xml:id="cdbach" sex="2">
 <persName>Catharina Dorothea Bach</persName>
</person>
<person xml:id="ghbach" sex="1">
 <persName>Gottfried Heinrich Bach</persName>
</person>
<!-- .... -->
<relationGrp type="children" subtype="first-marriage">
 <relation name="parent" active="#jsbach" pas-
sive="#cdbach"/>
<!-- .... -->
</relationGrp>
<relationGrp type="children" subtype="second-marriage">
 <relation name="parent" active="#jsbach" pas-
sive="#ghbach"/>
<!-- .... -->
</relationGrp>
```

Other kinds of entity

- <org>: a named collection of people regarded as a single unit, such as a business, institution, or tribe.
- <place>: a named location of any kind (including mythological and non-terrestrial places)
- These can be grouped in the same way (using <listOrg> or <listPlace>), and also have states, traits, and events.

Places

- Places can be identified solely in terms of geographical features or locations, e.g.

```
<place>
 <placeName>
  <geogFeat>mount</geogFeat>
  <geogName>Sinai</geogName>
 </placeName>
</place>
```

- More usually, they are identified in geo-political terms, using
  - adminstrative units such as <bloc>, <country>, <region>, <settlement>, <district>
  - physical location using <geo> and <offset>
- Note that all these things are traits — they may change over time

For example: Mayakovsky's birth place

```
<place xml:id="BGDT">
 <placeName xml:lang="ka">ბაღდათი</placeName>
 <placeName xml:lang="en">Baghdati</placeName>
 <placeName notAfter="1990" notBefore="1940">
Mayakovsky</placeName>
 <location type="geopolitical">
  <country>Georgia</country>
  <region type="geog">Imereti</region>
 </location>
 <location type="physical">
  <offset>West of</offset>
  <placeName>
   <geogFeat>River</geogFeat>
   <geogName>Khanistskali</geogName>
  </placeName>
  <geo>42.102298, 42.832947</geo>
 </location>
 <population when="2007">
  <p>4,700 people</p>
 </population>
</place>
```

Places can be nested (unlike people)

```
<place xml:id="LT">
 <country>Lithuania</country>
 <country xml:lang="lt">Lietuva</country>
 <place xml:id="LT-VN">
  <settlement>Vilnius</settlement>
 </place>
 <place xml:id="LT-KA">
  <settlement>Kaunas</settlement>
 </place>
</place>
```

Sources

Responsibility and uncertainty about the sources can be asserted by using attributes from the att.editLike class:

```
<org xml:id="MXY" type="tribe" resp="#herodotus">
 <orgName>The Maxyans</orgName>
 <country>Libya</country>
 <desc>According to Herodotus, they were a west Libyan
   tribe who said that they were descended from the men
   of Troy.</desc>
</org>
```

Dates and Periods

The support for dates in TEI P5 has concentrated on enabling greater use of international standards (W3C and ISO)

- `<date>` contains a date in any format
- `<time>` contains a phrase defining a time of day in any format

Example

```
<place xml:id="leipzig-univ">
 <placeName>University of Leipzig</placeName>
 <event type="foundation">
  <desc>The university was founded on
<date when="1409-12-02">December 2, 1409</date>.
  </desc>
 </event>
</place>
```

W3C Date Formats

Thanks to the mapping to W3C (att.datable.w3c) and ISO date formats, automatic processing and validation of expression of dates and times are now allowed

att.datable.w3c provides attributes for normalization of elements that contain datable events using the W3C datatypes

@when	supplies the value of a date or time in a standard form
@notBefore	specifies the earliest possible date for the event in standard form
@notAfter	specifies the latest possible date for the event in standard form
@from	indicates the starting point of the period in standard form
@to	indicates the ending point of the period in standard form

The W3C standard form for dates is YYYY-MM-DD.

Example

```
<place xml:id="leipzig-univ2">
 <placeName>University of Leipzig</placeName>
<!--…-->
 <event type="opening" notBefore="1409-09-09">
  <desc>The <foreign xml:lang="la">Alma mater
       Lipsiensis</foreign> opened in 1409, after it
    had been officially endorsed by Pope Alexander
    V in his Bull of Acknowledgment on
    (September 9 of that year).</desc>
 </event>
</place>
```

ISO Date Fomats

For some uses the subset of ISO 8601 which is used by the W3C might not be enough, so the TEI provides an optional att.datable.iso class to give the following attributes if needed:

- *@when-iso* the value of a date or time in a standard form
- *@notBefore-iso* the earliest possible date for the event
- *@notAfter-iso* the latest possible date for the event
- *@from-iso* the starting point of the period
- *@to-iso* the ending point of the period
- *@dur-iso* the length of this element in time

The ISO standard, for example, allows specifying dates and durations with a precision by omitting some digits to the left, while the W3C datatypes require in most cases conformance to a stricter precision.

Example

```
<p>He arrived <time when="12:00:00">around noon</time>.  He
arrived <time when-iso="12">around noon</time>.</p>
```

Time Periods and Relative Chronology

Time periods and relative chronology can also be defined.

```
<encodingDesc>
  <classDecl>
    <taxonomy xml:id="periods">
      <category xml:id="hellenistic">
        <catDesc>
          <ref
              target="http://www.wikipedia.com/wiki/Hellenistic">
Hellenistic</ref>.  Commonly treated as
<date notBefore="-0323" notAfter="-0031"/>.  </catDesc>
      </category>
<!--…-->
    </taxonomy>
  </classDecl>
</encodingDesc>
<!--…-->
<p> The city was built near a marble quarry which was extensively exploited in
the
<date period="#hellenistic">Hellenistic</date> and
<date period="#roman">Roman</date> periods.</p>
```

Handling primary sources in TEI XML

TEI @ Oxford

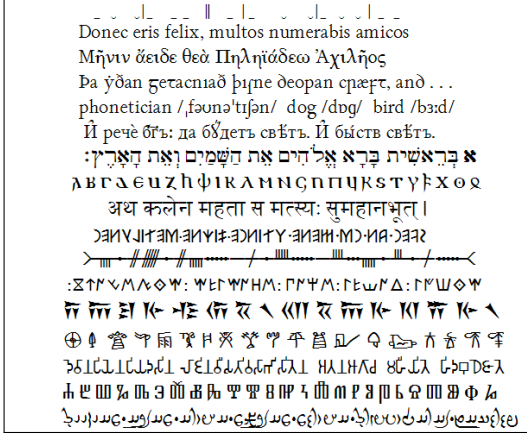
September 2008

Transcribable features

Transcription is a special kind of encoding, in which the aim is to represent all the important features of a primary source without prejudging too much about it... hence the term diplomatic transcript.
Here are some of the kinds of features concerned:

- letter forms
- page layout
- orthography
- word division
- punctuation
- abbreviations
- additions and deletions
- errors and omissions

Letter forms



Letter forms

- Unicode (ISO 10646) defines computer codepoints for most, though not all, of the abstract characters recognized by modern scholars when reading ancient sources.
- Different fonts realise those codepoints in different styles; however the underlying character remains the same.
- Data entry of Unicode characters can be
  - direct: some key combination or menu-selection generates the character æ for us
  - indirect, using a numeric character entity reference such as &#xE6
  - indirect using a mnemonic character entity reference such as æ (this requires every document to carry a DTD with it)

Non-Unicode characters

Nevertheless, sometimes Unicode is not enough…

- … if your character doesn't exist
- … if you want to distinguish letter forms that Unicode regards as identical e.g. for statistical analysis.

The <g> (gaiji) element stands for any non-Unicode character. Its content can be a local approximation to the desired letter (or nothing); its *@ref* attribute points to a definition for the required character or glyph.

```
<!-- in text --><g ref="#x123"/>
```

or

```
<g ref="#x123">x</g>
```

in header:

```
<char xml:id="x123">
<!-- character definition here -->
</char>
```

Structure and layout

- As elsewhere we distinguish 'structure' (the way the intellectual content of a work is logically organized) from 'layout' (the physical arrangement of the text on the page).
- The structural view is generally privileged over the layout view in TEI documents. Common practice is to mark <p>, <lg>, <l> (etc) elements, elements, as in printed texts, and to use empty 'milestone' tags for significant points in the physical layout, for example <pb/>, <cb/>, and <lb/>, for page-, column- and line-boundaries respectively.
- (The opposite practice is also feasible: one could imagine marking up a structural hierarchy of <gathering>, <leaf>, etc. with milestone elements to mark the points at which 'structural' components begin and end.)

Abbreviation

Abbreviations are highly characteristic of manuscript materials of all kinds. Western MSS traditionally distinguish:

Suspensions	the first letter or letters of the word are written, generally followed by a point, or other marker: for example e.g. for exempla gratia
Contractions	both first and last letters are written, generally with some other mark of abbreviation such as a superscript stroke, or, less commonly, a point or points: e.g. Mr. for Mister.
Brevigraphs	Special signs or tittels, such as the Tironian nota used for 'et', the letter p with a barred tail commonly used for per, the letter c with a circumflex used for cum (ĉ) etc
Superscripts	Superscript letters (vowels or consonants) are often used to indicate various kinds of contraction: e.g. w followed by superscript ch for which.

Encoding abbreviations (1)

TEI proposes two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion: <abbr> and <expan>
- abbreviatory signs or characters and the 'invisible' characters they imply: <am> and <ex>

The Old Icelandic word *hann* ('he') is usually written as a brevigraph, combining the letter h with a horizontal stroke representing nasalisation (Unicode character 0305, functionally similar to the modern tilde). It looks like this:

Encoding abbreviations (2)

Depending on editorial policy, we might represent this combination in any one of the following ways:

```
<abbr>h&#x305;</abbr>

<expan>hann</expan>

h<am>&#x305;</am>

h<ex>ann</ex>

<abbr>h<am>&#x305;</am>
</abbr>

<expan>h<ex>ann</ex>
</expan>
```

Encoding abbreviations (3)

We could also indicate multiple alternatives (at either level) by using the <choice> element

```
h<choice>
  <am>&#x305;</am>
  <ex>ann</ex>
</choice>
<choice>
  <abbr>h&#x305;</abbr>
  <expan>hann</expan>
</choice>
```

And much more besides...

Encoding abbreviations (3)

The *@type* attribute on <abbr> allows us to provide alternative renderings for the same markup in different contexts.

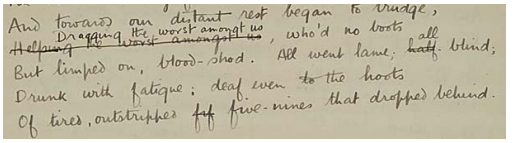
```
<choice>
  <abbr type="susp">k<am>&#x307;</am>
  </abbr>
  <expan>k<ex>onungr</ex>
  </expan>
</choice>
<choice>
  <abbr type="tittel">ml<am>&#x305;</am>i</abbr>
  <expan>m<ex>æl</ex>l<ex>t</ex>i</expan>
</choice>
```

*k(onungr) m**æll**ti*

As elsewhere, the *@resp* and *@cert* attributes can also be used to indicate who is responsible for an expansion, and the degree of certainty attached to it.

Additions, deletions, and substitutions

- <add> (addition) or <del> (deletion) are used for evident alterations in the source
- a combined addition and deletion may be marked using <subst> (substitution)



Additions, deletions, and substitutions

```
<l>And towards our distant rest began to trudge,</l>
<l>
  <subst>
    <del>Helping the worst amongst us</del>
    <add>Dragging the worst amongt us</add>
  </subst>, who'd no boots
</l>
<l>But limped on, blood-shod.  All went lame;
<subst>
    <del>half-</del>
    <add>all</add>
  </subst> blind;</l>
<l>Drunk with fatigue ; deaf even to the hoots</l>
<l>Of tired, outstripped <del>fif</del> five-nines that dropped behind.</l>
```

Corrections and emendations

The <sic> element can be used to indicate that the reading of the manuscript is erroneous or nonsensical, while <corr> (correction) can be used to provide what in the editor's opinion is the correct reading:

```
<sic>giorit</sic>
```

```
<corr>giorir</corr>
```

Alternatively, they may be combined within a <choice> element, thus allowing the possibility of providing multiple corrections:

```
<choice>
 <sic>giorit</sic>
 <corr cert="high">giorir</corr>
 <corr cert="low">gioret</corr>
</choice>
```

Supplied text

Sometimes, a transcript may need to include words not visibly present in the source:

- because the carrier has been damaged or is barely legible
- because of (assumed) scribal error

The <supplied> element is provided for use in either situations; the *@reason* attribute is used to distinguish them.

```
…Dragging the worst
among<supplied reason="omitted">s</supplied>t us…
```

Metadata for supplied text

Attributes *@resp* and *@cert* can be used here as elsewhere. A *@source* attribute is also available to indicate that another witness supports the reconstruction:

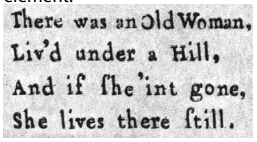
```
<p>ath þeir <supplied reason="omitted" source="AM02-152">mundu</supplied>
sundr ganga</p>
```

When missing text cannot be confidently reconstructed, the <gap> element should be used. Its *@reason* attribute explains the reason for the omission and its *@extent* attribute indicates its presumed size.

```
<gap reason="damage" extent="7cm"/>
```

Normalization

Source texts rarely use modern normalized orthography. For retrieval and other processing reasons, such information may be useful in a transcription. The <reg> (regularized) element is available used to mark a normalized form; the <orig> (original) element to indicate a non-standard spelling. These elements can optionally be grouped as alternatives using the <choice> element:


There was an Old Woman,
Liv'd under a Hill,
And if ſhe 'int gone,
She lives there ſtill.

Normalization example

```
<lg>
 <l>There was an Old Woman,</l>
 <l>
  <choice>
   <orig>Liv'd</orig>
   <reg>Lived</reg>
  </choice> under a hill,</l>
 <l>And if she <orig>'int</orig> gone,</l>
 <l>She lives there still.</l>
</lg>
```

Why are manuscript descriptions special?

- Manuscripts are *unique objects*, sometimes (though not always) of great cultural or political value
- Books, by contrast, exist in multiple copies, and can be described adequately by well-established and formalized bibliographic conventions.
- For manuscripts, there are several traditions, often descriptive or belle lettriste, and little consensus.

Similar concerns apply to other text-bearing objects.

Objectives

The TEI <msDesc> element is intended for several different kinds of applications:

- standalone database of library records (finding aid)
- discursive text collecting many records (catalogue raisonné)
- metadata component within a digital surrogate (electronic edition)
- tool for 'quantitative codicology'

Catalogue Raisonné

An <msDesc> can appear anywhere a <p> paragraph can

```
<div>
 <head>The Arnamagnæan Institute and its records</head>
 <p>Probably the finest collection of ……
 </p>
 <p>For example:  </p>
 <msDesc xml:id="AMI-1" xml:lang="en">
<!-- …-->
 </msDesc>
 <p>In the following manuscript….
 </p>
 <msDesc xml:id="AMI-2" xml:lang="en">
<!-- …-->
 </msDesc>
</div>
```

Digital edition

- metadata in the header
- transcription in the body, with links to
- images in a <facsimile> element

```
<TEI>
 <teiHeader>
<!-- … metadata describing the manuscript -->
<!-- includes a msDesc within the sourceDesc -->
 </teiHeader>
 <facsimile>
<!-- … metadata describing the digital images -->
 </facsimile>
 <text>
<!-- (optional) transcription of the manuscript -->
 </text>
</TEI>
```

Example minimal structure

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>[Title of manuscript]</title>
  </titleStmt>
  <publicationStmt>
   <distributor>[name of data provider]</distributor>
   <idno>[project-specific identifier]</idno>
  </publicationStmt>
  <sourceDesc>
   <msDesc xml:id="ex1" xml:lang="en">
<!-- [full manuscript description ]-->
   </msDesc>
  </sourceDesc>
 </fileDesc>
 <revisionDesc>
  <change when="2008-01-01">[revision information]</change>
 </revisionDesc>
</teiHeader>
```

Quantitative Codicology: is it possible?

Two conflicting desires:

- preserve (or perpetuate) existing descriptive prose
- reliable search, retrieval, and analysis of data

The <msDesc> tries, wherever possible, to have its cake and eat it.

Components of a manuscript description

We separate, and tag differently, aspects concerned with…

- identification
- intellectual content
- physical description
- history and curation
- … and other manuscript descriptions

msDesc structure

```
<msDesc xml:id="ex2" xml:lang="en">
 <msIdentifier>
<!-- Repository location, shelfmarks, etc. -->
 </msIdentifier>
 <msContents>
<!-- Structured description of MS contents -->
 </msContents>
 <physDesc>
<!-- Physical and codicological description -->
 </physDesc>
 <history>
<!-- Origin, provenance, acquisition, etc. -->
 </history>
 <additional>
<!-- Additional bibliographic and curatorial information,
and associated materials etc. -->
 </additional>
 <msPart>
<!--Composite manuscript details -->
 </msPart>
</msDesc>
```

<msIdentifier> is the only one that is required.

Simple example <msDesc>

```
<msDesc xml:id="ex3" xml:lang="en">
  <msIdentifier>
    <settlement>Oxford</settlement>
    <repository>Bodleian Library</repository>
    <idno>MS. Add. A. 61</idno>
    <altIdentifier type="other">
      <idno>28843</idno>
    </altIdentifier>
  </msIdentifier>
  <p>In Latin, on parchment: written in more than one hand of the 13th
     cent. in England: 7⅛ x 5⅝ in., i + 55 leaves, in double columns: with
     a few coloured capitals.</p>
  <p>'Hic incipit Bruitus Anglie,' the De
     origine et gestis Regum Angliae of Geoffrey of Monmouth (Galfridus
     Monumetensis: beg. 'Cum mecum multa &amp;amp; de multis.'</p>
  <p>On fol. 54v very faint is 'Iste liber est fratris guillelmi de
     buria de … Roberti ordinis fratrum Pred[icatorum],' 14th cent. (?):
     'hanauilla' is written at the foot of the page (15th cent.). Bought
     from the rev. W. D. Macray on March 17, 1863, for £1 10s.</p>
</msDesc>
```

Structured form of <msDesc> (1)

```
<msDesc xml:id="ex4" xml:lang="en">
  <msIdentifier>
    <settlement>Oxford</settlement>
    <repository>Bodleian Library</repository>
    <idno>MS. Add. A. 61</idno>
    <altIdentifier type="internal">
      <idno>28843</idno>
    </altIdentifier>
  </msIdentifier>
  <msContents>
    <msItem>
      <author xml:lang="en">Geoffrey of Monmouth</author>
      <author xml:lang="la">Galfridus Monumetensis</author>
      <title type="uniform" xml:lang="la">De origine et gestis Regum
Angliae</title>
      <rubric xml:lang="la">Hic incipit Bruitus Anglie</rubric>
      <incipit xml:lang="la">Cum mecum multa & de multis</incipit>
      <textLang mainLang="la">Latin</textLang>
    </msItem>
  </msContents>
<!-- … -->
</msDesc>
```

Structured form of <msDesc> (2)

```
<physDesc>
  <objectDesc form="codex">
    <supportDesc material="perg">
      <support>
        <p>Parchment.</p>
      </support>
      <extent>i + 55 leaves <dimensions scope="all" type="leaf" unit="in">
          <height>7 ⅛</height>
          <width>5 ⅝</width>
        </dimensions>
      </extent>
    </supportDesc>
    <layoutDesc>
      <layout columns="2">
        <p>In double columns.</p>
      </layout>
    </layoutDesc>
  </objectDesc>
  <handDesc>
    <p>Written in more than one hand.</p>
  </handDesc>
  <decoDesc>
    <p>With a few coloured capitals.</p>
  </decoDesc>
</physDesc>
```

Structured form of <msDesc> (2)

```
<history>
  <origin>
    <p>Written in <origPlace>England</origPlace> in the
<origDate notAfter="1300" notBefore="1200">13th cent.</origDate>
    </p>
  </origin>
  <provenance>
    <p>On fol. 54v very faint is <quote xml:lang="la">Iste liber est
          fratris guillelmi de buria de
      <gap reason="illegible"/> Roberti ordinis
          fratrum Pred<ex>icatorum</ex>
      </quote>, 14th cent. (?):
    <quote>hanauilla</quote> is written at the foot of
        the page (15th cent.).</p>
  </provenance>
  <acquisition>
    <p>Bought from the rev. <name type="person" key="MCRAYWD">W. D.
Macray</name> on
      <date when="1863-03-17">March 17,
          1863</date>, for £1 10s.</p>
  </acquisition>
</history>
```

Identification (1)

The <msIdentifier>
Traditional three part specification:

- place (<country>, <region>, <settlement>)
- repository (<institution>, <repository>)
- identifier (<collection>, <idno>)

```
<msIdentifier>
  <country>France</country>
  <settlement>Troyes</settlement>
  <repository>Bibliothèque Municipale</repository>
  <idno>50</idno>
</msIdentifier>
```

Identification (2)

Alternative or additional names can also be included:

```
<msIdentifier>
  <country>Danmark</country>
  <settlement>København</settlement>
  <repository> Det ArnamagnæanskeInstitut </repository>
  <idno>AM 45 fol.</idno>
  <msName  xml:lang="la">Codex Frisianus</msName>
  <msName  xml:lang="is">Fríssbók</msName>
</msIdentifier>
```

Intellectual Content

- May simply use paragraphs of text…
- … or a tree of <msItem> elements
- … optionally preceded by a prose summary

We can describe the content in general terms:

```
<msContents>
  <p>An extraordinary charivari of heroic deeds and
     improving tales, including an early version of
  <title>Guy of Warwick</title> and several hymns.
  </p>
</msContents>
```

or we can provide detail about each distinct item:

```
<msContents>
  <summary>An extraordinary charivari of heroic deeds,
     improving tales, and hymns</summary>
  <msItem>
<!-- details of Guy of Warwick here -->
  </msItem>
  <msItem>
<!-- other items here -->
  </msItem>
</msContents>
```

The <msItem> element

Manuscripts contain identifiable items, usually physically tied to a locus.

- <locus>, if present, must be given first
- then any of the following, in a specified order:
  - <author>, <respStmt>
  - <title>, <rubric>, <incipit>, <explicit>, <colophon>, <finalRubric>
  - <quote>, <textLang>, <decoNote>, <bibl>, <listBibl>, <note> …
  - … or nested <msItem>s

<msContents> with multiple <msItem>s

```
<msContents>
  <msItem  n="1">
    <locus>fols.  5r-7v</locus>
    <title>An ABC</title>
    <bibl>
      <title>IMEV</title>
      <biblScope  type="pages">239</biblScope>
    </bibl>
  </msItem>
  <msItem  n="2">
    <locus>fols.  7v-8v</locus>
    <title  xml:lang="fr">Lenvoy de Chaucer a Scogan</title>
    <bibl>
      <title>IMEV</title>
      <biblScope  type="pages">3747</biblScope>
    </bibl>
  </msItem>
<!-- … -->
  <msItem  n="6">
    <locus>fols.  14r-126v</locus>
    <title>Troilus and Criseyde</title>
    <note>Bk.  1: 71-Bk.  5: 1701,  with additional losses due to mutilation
        throughout</note>
  </msItem>
</msContents>
```

Physical Description

An artificial (but helpful) grouping of many distinct items.
You can simply supply paragraphs of prose, covering such topics as

- <objectDesc>: the physical carrier
- <handDesc>: what is carried on it
- <musicNotation>, <decoDesc>, <additions>
- <bindingDesc> and <sealDesc>
- <accMat>: accompanying material

Or, group your discussion within the specific elements mentioned above.
Similarly, within the specific elements, you can supply paragraphs of prose, or further specific elements.

The carrier 1

The <objectDesc> contains just paragraphs, or <supportDesc> and <layoutDesc>

```
<objectDesc form="codex">
  <supportDesc material="mixed">
    <p>Early modern <material>parchment</material> and
    <material>paper</material>. </p>
  </supportDesc>
  <layoutDesc>
    <layout columns="1" ruledLines="25 32"/>
  </layoutDesc>
</objectDesc>
```

The carrier 2

A more complex substructure with specific elements for <support>, <extent>, <foliation>, <collation>, <condition>.
Multiple layouts may also be specified:

```
<layoutDesc>
  <layout ruledLines="25" columns="1">
    <p>
      <locus from="1r-202v"/>
      <locus from="210r-212v"/>
        Between 25 and 32 ruled lines. </p>
  </layout>
  <layout ruledLines="34 50" columns="1">
    <p>
      <locus from="203r-209v"/>Between 34 and 50 ruled lines. </p>
  </layout>
</layoutDesc>
```

<handDesc> and <decoDesc>

- <handNote> (note on hand) describes a particular style or hand distinguished within a manuscript.
- <decoNote> contains a note describing either a decorative component of a manuscript, or a fairly homogenous class of such components.

<handDesc> examples

```
<handDesc hands="2">
  <p>The manuscript is written in two contemporary hands, otherwise
     unknown, but clearly those of practised scribes.  Hand I writes
     ff.  1r-22v and hand II ff.  23 and 24.  Some scholars, notably Verner
     Dahlerup and Hreinn Benediktsson, have argued for a third hand on
     f.  24, but the evidence for this is insubstantial. </p>
</handDesc>


<handDesc hands="3">
  <handNote xml:id="Eirsp-1" scope="minor" script="other">
    <p>The first part of the manuscript, <locus from="1v" to="72v: 4">fols
1v-72v: 4</locus>, is written in a
        practised Icelandic Gothic bookhand.  This hand is not
        found elsewhere. </p>
  </handNote>
  <handNote xml:id="Eirsp-2" scope="major" script="other">
    <p>The second part of the manuscript, <locus from="72v: 4" to="194v">fols
72v: 4-194</locus>, is
        written in a hand contemporary with the first; it can
        also be found in a fragment of <title>Knýtlinga
            saga</title>, <ref>AM 20b II fol. </ref>. </p>
  </handNote>
  <handNote xml:id="Eirsp-3" scope="minor" script="other">
    <p>The third hand has written the majority of the
        chapter headings.  This hand has been identified as the
        one also found in <ref>AM 221 fol. </ref>. </p>
  </handNote>
</handDesc>
```

<additions>

The <additions> element can be used to list or describe any additions to the manuscript, such as marginalia, scribblings, doodles, etc., which are considered to be of interest or importance.

```
<additions>
  <p>The text of this manuscript is not interpolated with
     sentences from Royal decrees promulgated in 1294, 1305
     and 1314.  In the margins, however, another somewhat
     later scribe has added the relevant paragraphs of
     these decrees, see pp.  8, 24, 44, 47 etc. </p>
  <p>As a humorous gesture the scribe in one opening of
     the manuscript, pp.  36 and 37, has prolonged the lower
     stems of one letter f and five letters þ and has them
     drizzle down the margin. </p>
</additions>
```

<accMat>

<accMat> (accompanying material) contains details of any significant additional material which may be closely associated with the manuscript being described, such as non-contemporaneous documents or fragments bound in with the manuscript at some earlier historical period.

```
<accMat> A copy of a tax form from 1947 is included in the
envelope with the letter.  It
is not catalogued separately.  </accMat>
```

History

- <origin>: where it all began
- <provenance>: everything in between
- <acquisition>: how you acquired it

<origin> is datable element and thus has attributes notBefore and notAfter

Example

```
<history>
 <origin>
  <p>Written in <origPlace>England</origPlace> in the
  <origDate notAfter="1300" notBefore="1200">13th
       cent. </origDate>
  </p>
 </origin>
 <provenance>
  <p>On fol. 54v very faint is <q>Iste liber
       est fratris guillelmi de buria de
<gap reason="illegible"/>
       Roberti ordinis fratrum Pred<expan>icatorum</expan>
   </q>,
     14th cent. (?): <q>hanauilla</q> is written at the
     foot of the page (15th cent.).</p>
 </provenance>
 <acquisition>
  <p>Bought from the rev. <name type="person">W. D.
Macray</name>on <date when="1863-03-17"> March 17,
1863</date>,
     for 1pound 10s.</p>
 </acquisition>
</history>
```

Additional information

- <adminInfo>: administrative information
- <surrogates>: information about other surrogates eg pictures
- <accMat>: accompanying material
- <listBibl>: bibliography

Administrative information

- record history
- availability
- custodial history
- miscellaneous remarks

```
<adminInfo>
 <recordHist>
  <source>
   <p>Information transcribed from <ref target="IMEV123">IMEV 123</ref>
   </p>
  </source>
 </recordHist>
 <custodialHist>
  <custEvent type="conservation" notBefore="1961-03" notAfter="1963-02">
   <p>Conserved between March 1961 and February 1963 at Birgitte Dalls
       Konserveringsværksted.</p>
  </custEvent>
  <custEvent type="photography" notBefore="1988-05-01" notAfter="1988-05-
30">
   <p>Photographed in May 1988 by AMI/FA.</p>
  </custEvent>
  <custEvent type="other" notBefore="1989-11-13" notAfter="1989-11-13">
   <p>Dispatched to Iceland 13 November 1989.</p>
  </custEvent>
 </custodialHist>
</adminInfo>
```

And finally

A <msDesc> can contain a nested <msDesc>, <msPart>, catering for a combination of two MSS, formerly distinct.

```
<msDesc xml:id="ex5" xml:lang="en">
 <msIdentifier>
  <msName xml:lang="la">Codex Suprasliensis</msName>
 </msIdentifier>
 <msPart>
  <altIdentifier type="partial">
   <settlement>Ljubljana</settlement>
   <repository>Narodna in univerzitetna knjiznica</repository>
   <idno>MS Kopitar 2</idno>
   <note>Contains ff. 10 to 42 only</note>
  </altIdentifier>
 </msPart>
 <msPart>
  <altIdentifier type="partial">
   <settlement>Warszawa</settlement>
   <repository>Biblioteka Narodowa</repository>
   <idno>BO 3.201</idno>
  </altIdentifier>
 </msPart>
 <msPart>
  <altIdentifier type="partial">
   <settlement>Sankt-Peterburg</settlement>
   <repository>Rossiiskaia natsional'naia biblioteka</repository>
   <idno>Q.p.I.72</idno>
  </altIdentifier>
 </msPart>
</msDesc>
```