XML and TEI in Practice

Lou Burnard
July 2001

What are people doing with XML and the TEI?

1. Text collections and digital libraries
2. Digital archives of primary source materials
3. Critical (and uncritical) editions
4. Language analysis and representation

Favourite kinds of material

- The stuff...
  - Transcripts, varyingly encoded, of original source documents
  - Page images
  - Associated metadata, sometimes in a database
- ... and its organization.
  - by author (the collected works of...)
  - by topic (readings in ...)
  - by association (an archive of ...)
  - by text (a digital edition/archive of ..)

Favourite storage and access methods

Storage:

- Many separate documents or fragments
- Virtual documents from a specialised repository

Access:

☞ primarily: web readability

☞ often: finding aids using sophisticated metadata

☞ occasionally: text analytic methods

There seems to be scope for R & D here...

Favourite delivery methods

- direct delivery of XML is still rare, but increasing
- specialised XML delivery tools (e.g. Dynaweb) are still widely used
- hand-crafted text retrieval tools are not uncommon
- on-the-fly conversion to HTML is not uncommon
- one-off conversion to HTML is frequent
- in some places the XML may even be inaccessible!
- dumbing-down software (e.g. to eBooks) may assume more importance

Digital archive examples

☞ Center for Electronic Text in the Law (University of Cincinnati, Law faculty)

☞ Thesaurus musicarum italicarum (Leiden University, Informatics)

☞ Victorian Women Writers Project (Indiana University Library)

☞ Toyota City Imaging project (Bodleian Library, Oxford)

Digital edition examples

- ☞ Piers Plowman Archive (IATH, University of Virginia)
- ☞ La Charette project (Princeton, Poitiers)
- ☞ Henrik Ibsen project (Oslo, Trondheim, Bergen)

Language corpus examples

- Multext East (Slovenian Academy et al)
- British National Corpus
- Silfide (Serveur Interactif pour la Langue Française, son Identité, sa Diffusion, son Étude)

etc... see the TEI Applications pages

FAQs

- is this text or is it data?
- which parts of this should be preserved and how?
- IPR: who owns this?
- accessibility: who will use this and for what?
- accountability: are we doing the Right Thing?
- etc.

... all of these have an effect on the technical solutions chosen.

Text and data

We love oppositions!

☞ structured vs unstructured

☞ metadata vs content

☞ interpretion vs transcription

But XML/TEI facilitates convergence

- text can be treated as data
- data can be treated as text
- all kinds of digital oject can be integrated

Techniques for convergence

- resource **management** (whether centralised or distributed) is crucial
- establish project-specific Guidelines and document them
- establish conventions for naming and identification of documents and document fragments
- establish which content will be subject to authority control and how
- use the right tools for the job

Authority control

- Not just about establishing preferred vocabulary
- Also a means of multiplying access points

```
<person id="p123">
  <name type="preferred">Alonso the Magnificent</name>
  <name type="other">Alonso de Cabesa de Vaca</name>
  <birth><date value="15891102">St Brigita’s Day,
     1589</date><placeName>Sevilla</placeName>
  </birth>
  <occupation>Tyrant</occupation>
  <figure entity="p123pic"/>
<!- etc etc ->
</person>
```

```
<p>.... and was owned by
<name role="owner" key="p123">Alonso the
Magnificent</name> ... </p>
```

Digital Preservation

- Scholarship implies a continuity of comprehension
  - it isn’t enough to preserve the data
  - we must also preserve its meaning
- XML/TEI encoding makes meaning explicit and independent of
  - software
  - hardware
  - usage
- ... within limits
- Other possible strategies include
  - emulation
  - accumulation
  - cryogenics

Text analysis: the next frontier

- Once we have made our digital surrogates, what then?
- Traditional activities:

  data discovery usually searching by external criteria

  data analysis usually searching by internal characteristics

  data synthesis usually by associating shared judgments
- What tools will help combine these approaches?

Three examples of TEI application software

- TEI web site
- SARA: a corpus analysis tool
- Phelix: an XML database system

TEI web site

The challenge: applying the TEI scheme to management, authoring, and maintenance of large documentary websites
Key features:

- a suitable DTD for authoring
- tools for conversion of legacy documents
- an effective change management system
- XSLT for rendering XML statically or dynamically

See http://www.oucs.ox.ac.uk/oucsdoc/allc.html for full discussion; proofs of the pudding are at http://www.oucs.ox.ac.uk and indeed http://www.tei-c.org

SARA

The challenge: support lexical analysis of very large amounts of richly encoded text
Key features:

- SGML aware search and retrieval of linguistic data
- Inverted file index of tags and content
- User-friendly windows client talking to special text retrieval engine
- Generalised to support any TEI conformant corpus

See http://www.hcu.ox.ac.uk/SARA for sample tutorials and downloads (also try Lampeter Corpus on your CD

Phelix

The challenge: support DBMS-style retrieval and management of richly encoded metadata fragments

Key features:

- Detailed set of TEI extensions
- XML tree is decomposed into relations representing XML structure, not its semantics
- XML fragments generated and rendered using XSLT
- Entirely web-based interface, held together with PHP

For proof of pudding, see http://janus.oucs.ox.ac.uk/master