Standards for digital encoding

Tomaž Erjavec

Institut für Informationsverarbeitung
Geisteswissenschaftliche Fakultät
Karl-Franzens-Universität Graz

Lecture 2: TEI and XSLT
10.11.2006

Lecturer

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana
http://nl.ijs.si/et/
tomaz.erjavec@ijs.si

Corpora and other language resources, standards, annotation, text-critical editions

Web page for this course:
http://nl.ijs.si/et/teach/graz06/standards/

Students: send emails!

Overview

1. Introduction
2. TEI background
3. TEI structure
4. Introduction to XSLT

Lab session:
writing a teiLite document, transforming to HTML with XSLT
What's in a text?

Upon Julia's Clothes

When as in silk my Julia goes,
Then, then (my thoughts) how sweetly flows
That liquefaction of her clothes.
Next, when I cast mine eyes and see
That brave Vibration each way free;
O how that glittering taketh me!

What's in a text (2)?

What's in a text (3)?
The ontology of a text

- Where is the text?
  - in the shape of letters and their layout?
  - in the original from which this copy derives?
  - in the ideas it brings forth? in their format, or their intentions?
- Texts are abstractions conjured up by readers.
- Markup encodes those 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

Styles of markup

- In the beginning there was *procedural* markup
  
  RED INK ON; print balance; RED INK OFF
- which being generalised became *descriptive* markup
  
  <balance type='overdrawn'>some numbers</balance>
- also known as encoding or annotation
descriptive markup allows for re-use of data
Some more definitions

- Markup makes explicit the distinctions we want to make when processing a string of bytes.
- Markup is a way of naming and characterizing the parts of a text in a formalized way.
- It’s (usually) more useful to markup what things mean than what they look like.

What does markup capture?

- Compare
  
  ```
  <head><title>Julia's Clothes</title></head>
  <p>When as in silks my <hi>Julia</hi> goes,<br />
  Then, then (me thinks) how sweetly flowes<br />
  That liquefaction of her clothes.<br />
  </p>
  ```

- and
  
  ```
  <s n="1" role="head">
  <w type="pp">Upon</w>
  <w type="np">Julia</w>'s Clothes
  </s>
  <s n="2" role="line">
  <w type="adv">Whenas</w>
  <w type="pp">in</w>
  <w type="nn2">silks</w>
  </s>
  ```

Likewise...

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

- and
  
  ```
  <Hwæt! we Gar-dena in gear-dagum</Hwæt!> 
  <lb>þeod-cyninga þrym gefrunon;<lb>
  <lb>hu ða æþelingas ellen fremedon;<lb>
  <lb>Oft Scyld Scefing sceaþena þreatum;<lb>
  ```
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
  - for different users

A useful mental exercise

- Imagine you are going to markup several thousand pages of complex material....
  - Which features are you going to markup?
  - Why are you choosing to markup this feature?
  - How reliably and consistently can you do this?
- Now, imagine your budget has been halved. Repeat the exercise!

What can the TEI do for you?

The TEI provides a framework for the definition of multiple schemas
- it defines and names several hundred useful textual distinctions
- it provides a set of modules that can be used to define schemas making those distinctions
- it provides a customization mechanism for modifying and combining those definitions with new ones using the same conceptual model
Where did the TEI come from?

- Originally, a research project within the humanities
- Sponsored by three professional associations
- Major influences
  - digital libraries and text collections
  - language corpora
  - scholarly datasets
- International consortium established June 1999
  (see http://www.tei-c.org/)

Goals of the TEI

- better interchange and integration of scholarly data
- support for all texts, in all languages, from all periods
- guidance for the perplexed: what to encode — hence, a user-driven codification of existing best practice
- assistance for the specialist: how to encode — hence, a loose framework into which unpredictable extensions can be fitted

These apparently incompatible goals result in a flexible and modular environment

TEI Guidelines

- A set of recommendations for text encoding, covering both generic text structures and some highly specific areas based on (but not limited by) existing practice
- A very large collection of element definitions with associated declarations for various schema languages
- A modular system for creating personalized schemas or DTDs from the foregoing

for the full picture see http://www.tei-c.org/Guidelines2/
Legacy of the TEI

- a way of looking at what 'text' really is
- a codification of current scholarly practice
- (crucially) a set of shared assumptions and priorities about the digital agenda:
  - focus on content and function (rather than presentation)
  - identify generic solutions (rather than application-specific ones)

Users of TEI

- Over 100 projects listed on the TEI project page
- Main areas:
  - digital libraries
  - text-critical editions
  - computer corpora
  - dictionaries

Versions of the Guidelines

- TEI P3 (1994) first public version:
  - SGML + book (1200pp) and soon also on the Web.
  - provides equal support for XML and SGML applications using the TEI schema;
  - error correction, while maintaining backward compatibility;
  - documents conforming to TEI P3 will not become illegal when processed with TEI P4.
- TEI P4 (2002):
  - implements more fundamental changes to the schemas, in line with current practice and identified problems, e.g. uses namespaces
  - no longer backward compatible (but a migration P4 to P5 XSLT exists)
  - Relax NG becomes the main schema language
- TEI P5 (2006...)
  - still somewhat fluid (details in schemas, Web presentation)
The general structure of TEI documents

- Burnard, Driscoll, Rahtz, TEI Training Course, Sofia 2005: Slides for TEI overview

TEI Lite

- TEI Lite is a particular parametrisation of TEI that “provides 90% of the elements needed for 90% of users”
- the TEI Lite P4 DTD can be found at http://www.tei-c.org/Lite/DTD/teilite.dtd

Lab session 1

- again, recipes
  - Bavarian-Style Pork Roast with Cabbage and Knödel and 2 others, e.g. from the Cabbage section
  - take teilite DTD and mark-up the documents according to TEI
  - make use of documentation provided at TEI Lite page
XSLT

Erjavec: Course at ESSLII 2005
Annotation of Language Resources,
Lecture II. XML-Related
Recommendations:
Formatting and Transforming XML
- ZVON tutorial
- W3schools
- ...