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!

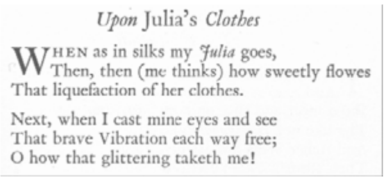
Overview

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

Lab session:

writing a teiLite document, trasforming to HTML with XSLT

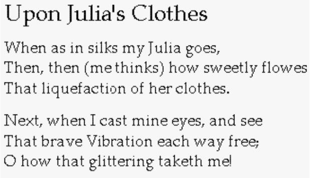
What’s in a text?



Upon Julia’s *Clothes*

WHEN as in silks my *Julia* goes,
Then, then (me thinks) how sweetly flowes
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!

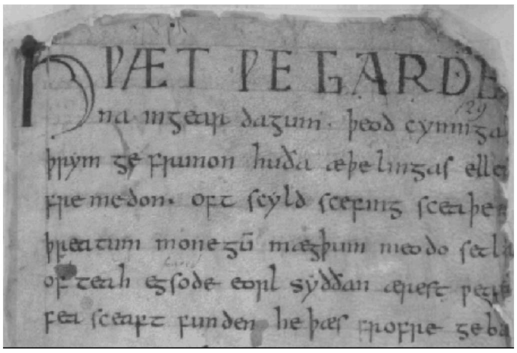


Upon Julia's Clothes

When as in silks my Julia goes,
Then, then (me thinks) how sweetly flowes
That liquefaction of her clothes.

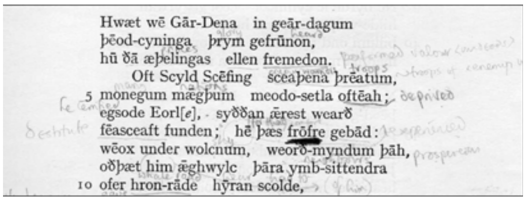
Next, when I cast mine eyes, and see
That brave Vibration each way free;
O how that glittering taketh me!

What’s in a text (2)?



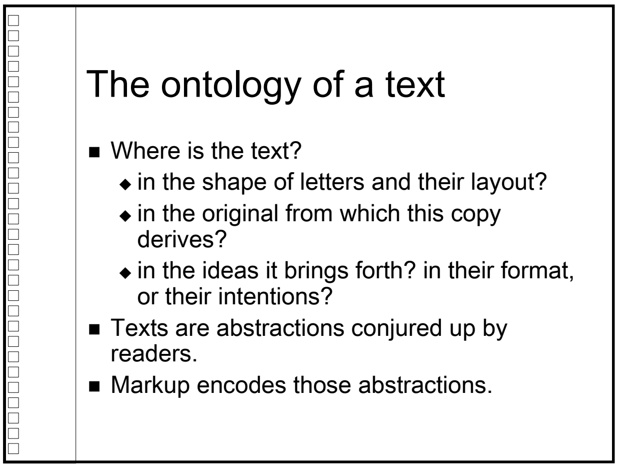
HWÆT WE GARDE
na ingear dagum. þeod cyninga
þrym ge frunon huða æþelingas elle
fre medon. oft scyld scefing sceaþe
þreatum monegū mægþum meodo setl
of teah egsode eorl syððan ærest wear
fea sceaft funden he þæs frofre geba

What’s in a text (3)?



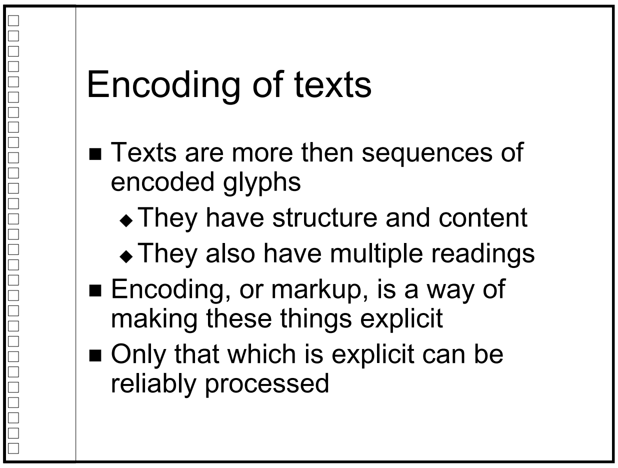
Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceaþena þrēatum,
monegum mǣgþum meodo-setla oftēah;
egsode Eorl[*e*], syððan ǣrest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him ǣghwylc þāra ymb-sittendra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

glory
heard
princes
performed valour (wonders)
troops
troops of enemies
many
nations
deprived
terrified
destitute
experienced
prosperous
whole road
had to
(of him)



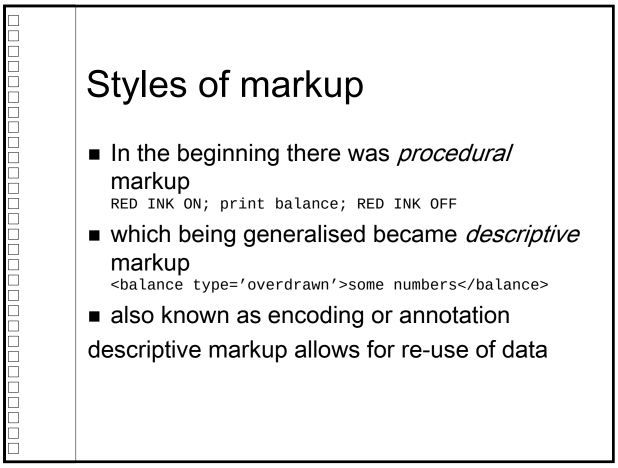
The ontology of a text

- 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

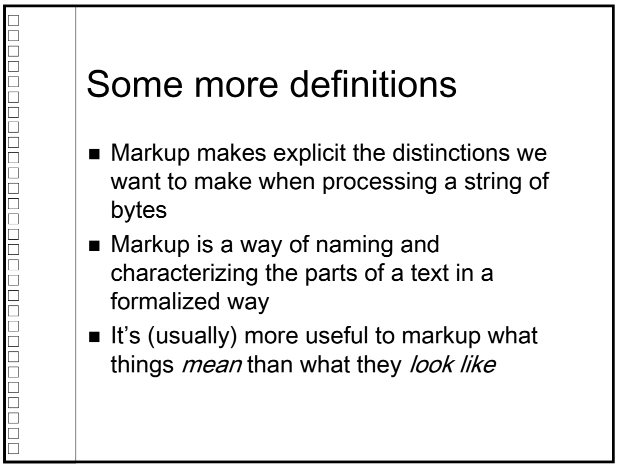
- Texts are more then sequences of encoded glyphs
  - They have structure and content
  - They also have multiple readings
- Encoding, or markup, is a way of making these things explicit
- Only that which is explicit can be reliably processed



Styles of markup

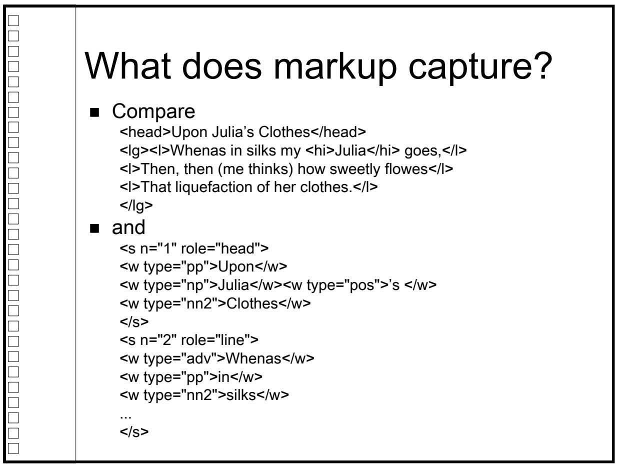
- In the beginning there was *procedural* markup
  `RED INK ON; print balance; RED INK OFF`
- which being generalised became *descriptive* markup
  `<balance type=’overdrawn’>some numbers</balance>`
- also known as encoding or annotation

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*



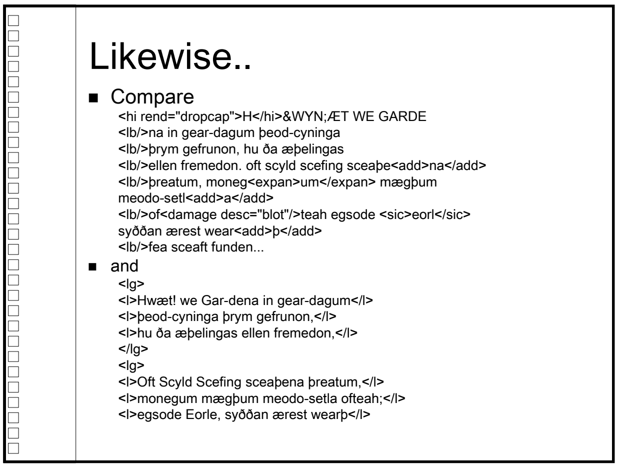
What does markup capture?

- Compare

```
<head>Upon Julia’s Clothes</head>
<lg><l>Whenas in silks my <hi>Julia</hi> goes,</l>
<l>Then, then (me thinks) how sweetly flowes</l>
<l>That liquefaction of her clothes.</l>
</lg>
```

- and

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



Likewise..

- Compare

```
<hi rend="dropcap">H</hi>&WYN;ÆT WE GARDE
<lb/>na in gear-dagum þeod-cyninga
<lb/>þrym gefrunon, hu ða æþelingas
<lb/>ellen fremedon. oft scyld scefing sceaþe<add>na</add>
<lb/>þreatum, moneg<expan>um</expan> mægþum
meodo-setl<add>a</add>
<lb/>of<damage desc="blot"/>teah egsode <sic>eorl</sic>
syððan ærest wear<add>þ</add>
<lb/>fea sceaft funden...
```

- 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>
```

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

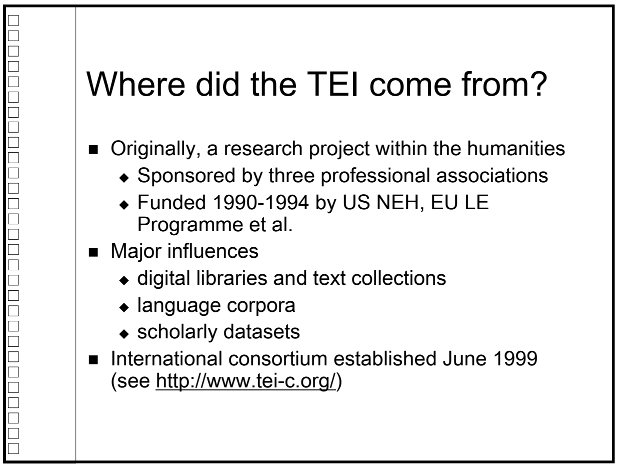
A useful mental exercise

- 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?

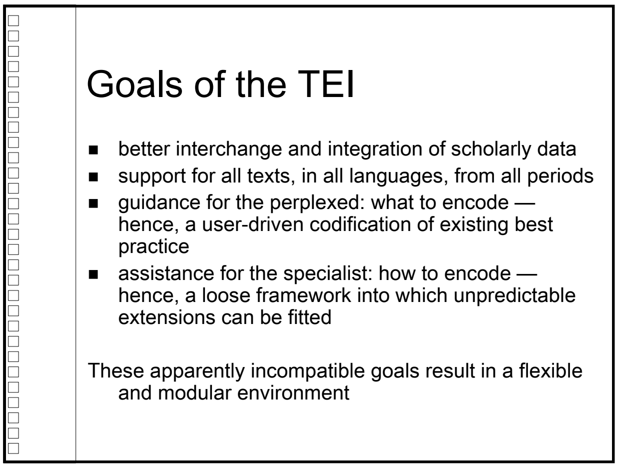
The TEI provides a framework for the definition of multiple schemas

- 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?

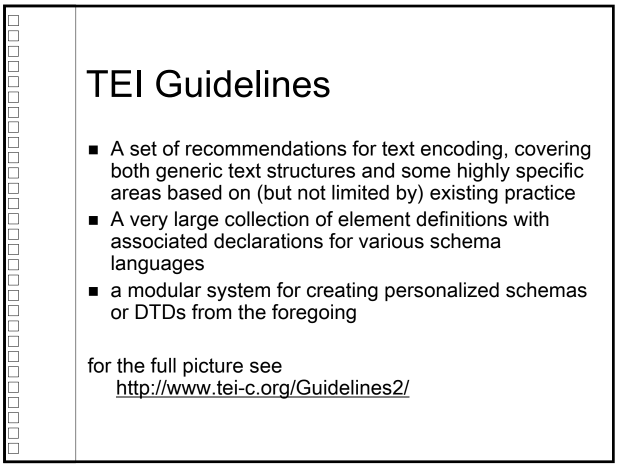
- Originally, a research project within the humanities
  - Sponsored by three professional associations
  - Funded 1990-1994 by US NEH, EU LE Programme et al.
- Major influences
  - digital libraries and text collections
  - language corpora
  - scholarly datasets
- International consortium established 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

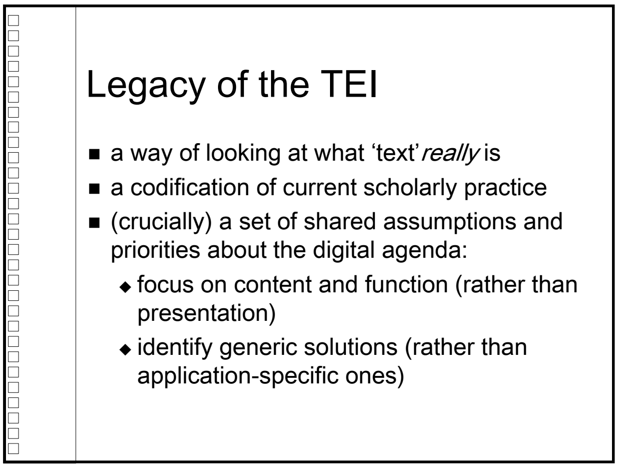
These apparently incompatible goals result in a flexible and modular environment



TEI Guidelines

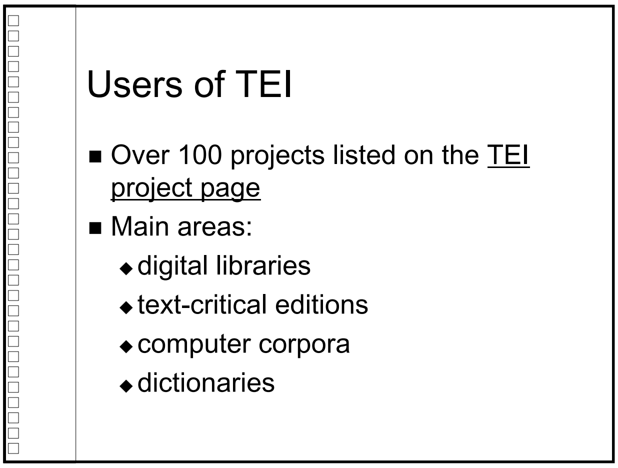
- A set of recommendations for text encoding, covering both generic text structures and some highly specific areas based on (but not limited by) existing practice
- A very large collection of element definitions with associated declarations for various schema languages
- a modular system for creating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

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



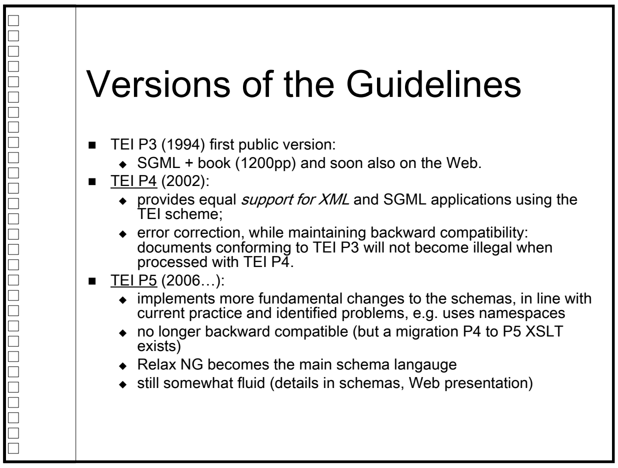
Legacy of the TEI

- 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.
- TEI P4 (2002):
  - provides equal *support for XML* and SGML applications using the TEI scheme;
  - error correction, while maintaining backward compatibility: documents conforming to TEI P3 will not become illegal when processed with TEI P4.
- TEI P5 (2006…):
  - 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 langauge
  - still somewhat fluid (details in schemas, Web presentation)

The general structure of TEI documents

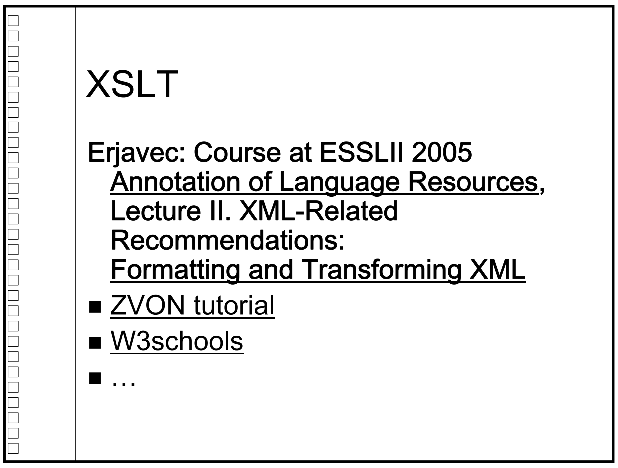
- Burnard, Driscoll, Rahtz, TEI Training Course, Sofia 2005:
  Slides for TEI overview

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
- …