Standards for language encoding: TEI

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute

ESSLLI 2011

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: a user-driven codification of existing best practice
- Assistance for the specialist — how to encode: a loose framework into which unpredictable extensions can be fitted

- Flexible and modular environment
- Also large and complex

What can the TEI do for you?

The TEI provides a framework for the definition of multiple (XML) 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
- Major influences
  - digital libraries and text collections
  - language corpora
  - scholarly datasets
- International consortium established June 1999 (see http://www.tei-c.org/)

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 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):
  - correction of errata
  - backward compatibility:
    provides *equal support for XML* and SGML
- TEI P5 (2006…):
  - more fundamental changes, in line with current practice and identified problems, e.g. uses namespaces
  - no longer backward compatible
  - Relax NG becomes the main schema language
  - regular new releases with small improvements

The general structure of TEI documents

Burnard, Driscoll, Rahtz, TEI Training Course, Sofia 2005:

Slides for TEI overview

TEI Community

- Yearly member meetings and conferences
- Journal of the Text Encoding Initiative
- Occasional tutorials (Oxford, Virginia)
- tei-l mailing list
- Web site
  - Guidelines, Tutorials
  - SIGs, WGs
  - Tools:
    - Roma
    - XSLT stylesheets
  - Quite different from ISO…

Roma

- „TEI provide a framework for the definition of multiple (XML) schemas“ How does it do this?
- „Literate programming“: TEI Guidelines are themselves written in TEI, using the ODD module
- The ODD language is, essentially, a schema language uses RELAX NG for the actual element / attribute definitions
- An ODD processor is able to produce RELAX, W3C or DTD schemas (+ documentation) from an ODD schema.
- TEI allows for changes and additions to the predefined options
- On-line ODD processor: Roma
- http://www.tei-c.org/Roma/
- Tutorial: http://tei.oucs.ox.ac.uk/Oxford/2011-07-oxford/presentations/tei-02-customising.pdf