TBX goes TEI

Laurent Romary, Inria & HUB
Implementing a TBX based extension for the TEI guidelines

• Addressing a new (?) user community: digital humanists
  – Importance of primary sources and construction of secondary digital objects (annotations, indexes, etc.)

• Taking the best from the onomasiological work in the last 40 years
  – Avoiding simplistic representations such as SKOS and thesaurus standards

• Fostering more convergence in standardization
  – Favoring reuse of components from various standardization worlds
The humanities have been going digital since half a century

- Roberto Busa (*Index Thomisticus*, 1946-...), Antonio Zampolli, ACH-ALLC
  - Linguistic, literature, archeology, etc.
  - More and more difficult to work without digital sources
- Dealing with a variety of documents
  - Prose, transcription of speech, drama, manuscripts, ... dictionaries
- Experience embodied in the Text Encoding Initiative (TEI) consortium
  - Very active community (conference, journal)
  - Efficient standardization process (technical council; 2 releases a year; deprecation mechanisms)
  - More than 500 elements
    - Embedded culture of customization
  - Strong specification platform
    - ODD: one document does it all
    - Literate programming: schema + documentation
- Note: quite a large community of XML aware people there
TEI and (semasiological) lexical data

• Seminal work from N. Ide and J. Véronis
  – Modular and (too?) flexible Dictionary chapter in the TEI
    • A lot of core components for orthographical, grammatical and semantic representations
  – Strongly inspired the work on LMF
  – TEI as a possible serialization of the core components of LMF (see “TEI and LMF crosswalks”)
    • Customization platform makes it easy to expand to deal with additional modules
  – Large community of users
    • See http://www.tei-c.org/Activities/Projects/
TEI and onomasiological representations

• A missed opportunity?
  – ISO 6156:1987 (Mater)
  – 1989: Setting up the TEI
    • Specific chapter of the TEI guidelines dedicated to the representation of terminological data
    • The SGML-based representation integrated in the TEI framework remained there until the P4 edition
  – ISO 12200 (Martif): 1999, improves the TEI proposal (bracketing), but breaks the link to the TEI by going ISO
    • document structure strongly inspired from the TEI (e.g. the header-text organisation; entries embedded within a <text> and <body> hierarchy);
    • reaching out to the translation and localisation industry
  – 1999-2003: Abstracting away
    • Basic for the specification of a variety of terminological formats
  – ISO 30042:2008 TBX (TermBase eXchange), after work carried out in LISA
  – Current: TBX-Basic, TBX-Min ...

• Let us make another member in the family
  – Assessing histocompatibility
Tissue typing

• Host: the TEI document structure
  – Terminological entries can occur at many places
    • Specific section, inline, between other TEI elements
  – As far as building up a terminological database in TEI
    • E.g. recording bilingual philosophical vocabulary from Wittgenstein’s works
  – Keeping all inline annotation facilities for textual fields
    • Names, dates, foreign expressions, notes, pointers, feature structures...
  – Improved documentation with the rich TEI header
    • Important from a scholarly perspective

• Graft: a TBX-like terminological entry
  – Structural skeleton
    • Inspired from TBX-Basic (DCA style)
    • Note that TBX already has an ODD spec!
  – Data categories
    • Initially reduced to a very small number of meaningful categories for a DH scenario
    • subjectField, definition, source, partOfSpeech, grammaticalGender, etc.
    • In particular: no project management data categories
The transplant process

Harvest and adapt

Insert wherever <entry> can occur
The role of the TEI class system

• Attribute classes
  – Group together attributes occurring within some elements in a regular manner
  – Global attributes, pointing, typing, dating etc.

• Model classes
  – Group together elements with a similar syntactic and/or semantic behavior
  – E.g. bibliographical objects
TBX in the ODD architecture
Surgery report

• Ensuring the graft by means of namespaces
  – Issue: no TBX namespace => we devised one
• Incompatible tissues
  – Attributes
    • att.global attribute class: @xml:id, @xml:lang, @xml:base, @xml:space
    • @target => att.pointing: making ID/IDREF be URI
  – Outdated element
    • <tbx:xref> (cf. URI mechanism)
  – TEI elements in their own namespace
    • <tei:term>
    • <tei:hi>: bringing the semantic back on tracks
    • <tei:ref>, <tei:ptr>, <tei:note>
• Second life
  – Rich textual content model
Industrie mécanique

endloser Riemen ...

De Coster, Wörterbuch ...

wird zum Antrieb der Lichtmaschine, des Ventilators ...

Keilriemen

De Coster, ...

…
A chimera?

• TBX@TEI is not a proper subset of TBX
  – Document structure
  – Changes in the content model of `<termEntry>`

• Still:
  – Can be used to generated TBX compatible data univocally
  – Is probably the most optimal way the get DH people to be acquainted with good terminological practices

• And the customization can be customized!
  – Getting rid of unwanted TEI objects
  – Providing description of more complex data categories
Next steps

- From a TEI point of view
  - Various options: new chapter, recommended customization
    - Available ODD file and test data
  - Documentation of data categories by means of DCR attributes
- From a TBX point of view
  - Can TBX@TEI be an acceptable member (a DH cousin)
  - Taking up some of the proposed (small) changes
- For the academic community and beyond
  - Maintenance of a family of formats at the service of various communities
  - Need to think of a global ODD-based strategy (library of components)