

Metadata as a Strategy for Domain-Specific Content Modelling

Tomaž Erjavec

<http://nl.ijs.si/et/>

Department of Knowledge Technologies

Jožef Stefan Institute

Ljubljana

Slovenia

Perspectives on Metadata: Digital Edition & Preservation

12.11-13.11.2009, Vienna

Overview of the talk

- Text Encoding Initiative Guidelines
- Metadata in TEI
- Examples
- Mapping TEI metadata
- Conclusions

Text Encoding Initiative


- TEI consortium develops and maintains a standard for the representation of texts in digital form
- Main deliverable: TEI Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.
- TEI Guidelines are used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation.
- The Consortium also provides supporting resources: materials for learning TEI, question-friendly mailing list, TEI-related publications, and software developed for or adapted to the TEI.

TEI: Text Encoding Initiative - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.tei-c.org/index.xml

Mail :: Welcome to Ho... Welcome to RIVET - Pl... CORDIS : ICT : Progr... The CQP Query Langu... Most Visited CiteULike Popup Post



< Text Encoding Initiative >

Home Guidelines Activities Tools Membership Support About News Online Store

Home Entire site

TEI Events

TEI Conference and Members Meeting 2009 (Ann Arbor MI, USA. 11-15 November). [[read more](#)]

TEI Webstore Opens. [[read more](#)]

Print copies of the P5 Guidelines now available. [[read more](#)]

News

21 August 2009: TEI Webstore Opens. [[read more](#)]

22 June 2009: Nominations for Board and Council Close July 1 [[read more](#)]

31 May 2009: Call for Participation: Advanced Seminars in TEI Encoding [[read more](#)]

31 May 2009: TEI Cosponsored Meeting: Balisage 2009 Program Announced. [[read more](#)]

TEI: Text Encoding Initiative

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including [resources for learning TEI](#), information on [projects using the TEI](#), TEI-related [publications](#), and [software](#) developed for or adapted to the TEI.

The TEI Consortium is a non-profit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors.

Want to become active in the TEI community? [Become a TEI Member](#), join a [Special Interest Group](#), sign up for the [TEI-L mailing list](#), and come to our [annual meetings](#).

Done

TEI Guidelines

- TEI Guidelines for Electronic Text Encoding and Interchange: schema specifications and accompanying documentation (1200 pp)
 - ◆ TEI P3 (1994): SGML DTD
 - ◆ TEI P4 (2002): errata, support for SGML + XML, backward compatible with P3
 - ◆ TEI P5 (2007): XML only, ISO RelaxNG as main schema language, not backward compatible (but migration supported)

Structure of the Guidelines

- Guidelines do not define one schema, but allow for creation of a TEI schema by making a TEI parametrisation:
 - ◆ combining various modules
 - ◆ making controlled deletions / additions / changes
- TEI Guidelines are written in XML, using the TEI ODD module (text + schema)
- Use of on-line service to make schema + documentatin from a TEI parametrisation

TEI meta-data

- The <teiHeader> is a top level obligatory TEI element
- Rich structure:
 - ◆ *file description* <fileDesc>
full bibliographical description of the computer document
 - ◆ *encoding description* <encodingDesc>
relationship between an electronic text and its source(s)
 - ◆ *text profile* <profileDesc>
classificatory and contextual information about the text
 - ◆ *revision history* <revisionDesc>
changes made during the development of the electronic text

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" version="5" xml:lang="slv">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Slovenski biografski leksikon</title>
        <respStmt> [4 lines]
        <respStmt> [4 lines]
        <respStmt> [4 lines]
      </titleStmt>
      <editionStmt> [2 lines]
      <publicationStmt>
        <publisher>Slovenska akademija znanosti in umetnosti</publisher>
        <publisher>Znanstvenoraziskovalni center SAZU</publisher>
        <pubPlace>Ljubljana</pubPlace>
        <date>2009</date>
        <idno type="ISBN">ISBN 978-961-268-001-5</idno>
        <availability> [4 lines]
        <pubPlace> [3 lines]
      </publicationStmt>
      <sourceDesc>
        <biblStruct>
          <monogr> [11 lines]
        </biblStruct>
      </sourceDesc>
    </fileDesc>
```


Case study I.

Digital library TEI metadata

- eZISS digital library, <http://nl.ijs.si/e-zrc/>
- TEI encoded critical editions of Slovenian literature, with HTML view
- Interconnected facsimiles, transcriptions, scholarly commentary, in some cases including audiovisual recordings
- Creative Commons licence
- TEI headers of editions include varied information about the edition

Manuscript descriptions

- Škofja loka Passion Play: oldest performance text in Slovene (~1715)
- eZISS edition contains facsimile of the play as well as related facsimiles
- TEI header gives the description of the manuscripts:
 - ◆ identifier, contents, physical description, history, ...

```

- <sourceDesc>
  - <msDesc xml:id="sp">
    - <msIdentifier>
      <country>Slovenija</country>
      <settlement>Škofja Loka</settlement>
      <repository>Kapucinski samostan v Škofji Loki</repository>
    </msIdentifier>
  - <msContents defective="false">
    - <summary>
      Na 51 folijih (šest nepopisanih) je napisano besedilo
      <hi rend="italic">Škofjeloškega pasijona</hi>
      v latinični lepopisni kurzivi in nemški kurzivni pisavi Kurrent. Rokopis sestavlja osem leg, vezan je v
      platnice nekdanjega urbarja Loškega gospostva s konca 17. stoletja. Rokopis je bil vezan kmalu
      potem, ko je bil okrog leta 1727 napisan.
    </summary>
  - <msItemStruct defective="false">
    <author xml:lang="slv">Oče Romuald Štandreški</author>
    <author xml:lang="lat">Romualdus a S. Andrea</author>
    - <title xml:lang="lat" type="editorial">
      Instructio pro Processione Locopolitana in die Parasceve Dni
    </title>
    <title xml:lang="slv" type="editorial">Škofjeloški pasijon</title>
  - <incipit xml:lang="lat" defective="false">
    Ad repetitam Instantiam, et enixas preces
    - <abbr type="suspension">
      Jll
      <hi rend="sup">mi</hi>
    </abbr>

```

Character descriptions

- Freising Manuscripts are the earliest (~A.D 1000) preserved writings in Slovenian as well as the earliest Slavic texts, written in the Latin alphabet.
- Digital edition contains several (diplomatic, critical phonetic) transcriptions, 6 translations, studies and commentaries, glossary, bibliography, ...
- Some transcriptions use special characters, not available in Unicode
- A special font (using PUA) exists to exactly render the special characters
- How to enable viewing FM with standard or special fonts?

TEI solution

Text <g>:

```
popravliti, ceravno ne zara sprema. (op. k. w. hater/</note/>/line>
<line n="21" id="bsDT.1.021">ba vuelaica. Bofe gozpodi miloštivi . tebe ze mil</line>
<line n="22" id="bsDT.1.022">tuori<g corresp="zrcolaEB81"/>. od. žih poftenih greh.
<line n="23" id="bsDT.1.023">I. vuénfih. í minfih. Efe iezem ztvoril. teh ze.<g corr
<line n="24" id="bsDT.1.024">miltuori<g corresp="zrcolaEB81"/>. I'. žuetei marii. I'.
```

TEI header <charDesc>:

```
<charDesc>
  <desc>Opis znakov vsebuje uporabljene znake s podroc&#269;ja zasebne rabe uniko
  <char id="zrcolaE137">
    <charName>CYRILLIC SMALL LETTER YERU WITH HOOK</charName>
    <charProp>
      <localName>font</localName>
      <value>ZRCola</value>
    </charProp>
    <charProp>
      <localName>mapping</localName>
      <value>lossy</value>
    </charProp>
    <mapping type="PUA">&#xE137;</mapping>
    <mapping type="standard">&#x044B;<!--CYRILLIC SMALL LETTER YERU--></mapping>
  </char>
```

Case study II.

Language Corpora

- Annotated text (or speech) collections for linguistic investigations or datasets for HLT research
- Contain descriptions of included texts:
 - ◆ Bibliographic
 - ◆ Taxonomic
 - ◆ Linguistic annotation specifications
- Linguistic annotations:
 - ◆ Lemma and PoS
 - ◆ Syntactic structures (treebanks)
 - ◆ Named entities, word-senses, anaphora, ...

FidaPLUS

- Reference corpus of Slovene
<http://www.fidaplus.net/>
- 600 mil. words, 20,000 texts of modern day written Slovene
- Each text annotated with text type
- Each word annotated with lemma and morphosyntactic description
- Encoded in ~TEI

Taxonomy

- 3 taxonomies

- Medium

- Text type

- Proofread

```
<taxonomy>
  <category xml:id="Ft.P">
    <catDesc>
      <term xml:lang="sl">prenosnik</term>
      <term xml:lang="en">medium</term>
    </catDesc>
  </category>
  <category xml:id="Ft.P.G">
    <catDesc>
      <term xml:lang="sl">govorni</term>
      <term xml:lang="en">spoken</term>
    </catDesc>
  </category>
  <category xml:id="Ft.P.E">
    <catDesc>
      <term xml:lang="sl">elektronski</term>
      <term xml:lang="en">electronic</term>
    </catDesc>
  </category>
  <category xml:id="Ft.P.P">
    <catDesc>
      <term xml:lang="sl">pisni</term>
      <term xml:lang="en">written</term>
    </catDesc>
  </category>
  <category xml:id="Ft.P.P.O">
    <catDesc>
      <term xml:lang="sl">objavljeno</term>
      <term xml:lang="en">published</term>
    </catDesc>
  </category>
  <category xml:id="Ft.P.P.O.K">
    <catDesc>
      <term xml:lang="sl">knjižno</term>
      <term xml:lang="en">book</term>
    </catDesc>
  </category>
</taxonomy>
```


Morphosyntactic annotation

Text msd attribute:

```
<c xml:id="F0028708.9.1.1">"</c>
<w xml:id="F0028708.9.1.2" lemma="tisti" msd="Zk-mer">Tistega</w> <S/>
<w xml:id="F0028708.9.1.3" lemma="večer" msd="Somer">večera</w> <S/>
<w xml:id="F0028708.9.1.4" lemma="biti" msd="Gp-spe-n">sem</w> <S/>
```

TEI header feature-structure library:

```
<fvLib>
<fs xml:id="Ncmsn" xml:lang="en" feats="#NO. #N1.c #N2.m #N3.s #N4.n" />
<fs xml:id="Ncmsg" xml:lang="en" feats="#NO. #N1.c #N2.m #N3.s #N4.g" />
<fs xml:id="Ncmsd" xml:lang="en" feats="#NO. #N1.c #N2.m #N3.s #N4.d" />
<fs xml:id="Ncmsan" xml:lang="en" feats="#NO. #N1.c #N2.m #N3.s #N4.a #N5.n" />
<fs xml:id="Ncmsay" xml:lang="en" feats="#NO. #N1.c #N2.m #N3.s #N4.a #N5.y" />
```

TEI header feature library:

```
<fLib>
<!--1. Noun (N)-->
<f name="PoS" xml:id="NO." xml:lang="en"><symbol value="Noun" /></f>

<f name="Type" xml:id="N1.c" xml:lang="en"><symbol value="common" /></f>
<f name="Type" xml:id="N1.p" xml:lang="en"><symbol value="proper" /></f>

<f name="Gender" xml:id="N2.m" xml:lang="en"><symbol value="masculine" /></f>
<f name="Gender" xml:id="N2.f" xml:lang="en"><symbol value="feminine" /></f>
<f name="Gender" xml:id="N2.n" xml:lang="en"><symbol value="neuter" /></f>
```

III. TEI metadata at work

1. HTML rendering
2. Concordancer attributes
3. Repository search

Viewing the header

- Header can be multilingual (@xml:lang)
- Localisation: TEI elements are given multilingual glosses
- A simple XSLT to convert to HTML
- XSLT displays only chosen language
- TEI elements connected to their definitions

TEI Header

§file description	§title statement	§title	Morphosyntactically tagged corpus jos1M (Slovene tags)		
		§principal researcher	§name id = ET	Tomaž Erjavec	
			§address	Department of Knowledge Technologies Jožef Stefan Institute Jamova cesta 39 SI-1000 Ljubljana Slovenia	
		§statement of responsibility	§name id = SIMON	Simon Krek	
			§responsibility	Linguistic annotation supervision.	
§statement of responsibility	§name id = ANTTA	Anita Drvoderič			
	§responsibility	Morphosyntactic annotation.			
§statement of responsibility	§name id = DEJAN	Dejan Hribar			
	§responsibility	Morphosyntactic annotation.			
§statement of responsibility	§name id = KAJA	Kaja Hadalin			
	§responsibility				

kolofon TEI

§ opis datoteke

§ zapis naslova

§ naslov

Oblikoslovno označeni korpus jos1M (slovenske oznake)

§ nosilec raziskave

§ ime

identifikator = ET

Tomaž Erjavec

§ naslov

Odsek za tehnologije znanja
Institut "Jožef Stefan"
Jamova cesta 39
1000 Ljubljana

§ zapis o odgovornosti

§ ime

identifikator = SIMON

Simon Krek

§ odgovornost

Vodenje oblikoslovnega označevanja.

§ zapis o odgovornosti

§ ime

identifikator = ANITA

Anita Drvoderič

§ odgovornost

Oblikoslovno označevanje.

§ zapis o odgovornosti

§ ime

identifikator = DEJAN

Dejan Hribar

§ odgovornost

Oblikoslovno označevanje.

§ zapis o odgovornosti

§ ime

identifikator = KAJA

Kaja Hadalin

§ odgovornost

Oblikoslovno označevanje.



<titleStmt>

[Home](#) | [Table of contents](#)
[C Elements](#)

<titleStmt> (title statement) groups information about the title of a work and those responsible for its intellectual content. [2.2.1 The Title Statement](#) [2.2 The File Description](#)

Module	header — 2 The TEI Header
Used by	biblFull fileDesc
May contain	core: author editor respStmt title header: funder principal sponsor
Declaration	<div style="border: 1px solid #ccc; padding: 5px;"> <pre>element titleStmt (att.global.attributes, (title+, model.respLike*))</pre> </div> <div style="text-align: right; font-size: small;">Copy to XML format</div>
Example	<div style="border: 1px solid #ccc; padding: 5px;"> <pre><titleStmt> <title>Capgrave's Life of St. John Norbert: a machine-readable transcription</title> <respStmt> <resp>compiled by</resp> <name>P. J. Lucas</name> </respStmt> </titleStmt></pre> </div> <div style="text-align: right; font-size: small;">Show all</div>

[\[English\]](#) [\[Deutsch\]](#) [\[Español\]](#) [\[Italiano\]](#) [\[Français\]](#) [\[日本語\]](#) [\[한국어\]](#) [\[中文\]](#)
[TEI Consortium](#) | [Feedback](#)

Copyright TEI Consortium 2007. Licensed under the GPL. Copying and redistribution is permitted and encouraged. Version 1.5.0. Last updated on November 8th 2009. This page generated on 2009-11-09T12:11:13Z.

Concordance attributes

- Some concordancers can also search & display arbitrary positional attributes of tokens
- E.g. lemmea and tags such as “Ncms”
- With the feature library, it is simple to decompose such tags into features, e.g. PoS=Noun, Type=common, Gender=masculine, Number=singular
- This makes it possible to search and display not only tags but also features

Query: IKORPUS; [pos="pridevnik"] [pos="pridevnik"] [pos="pridevnik"] [pos="samostalnik"]

N°	Hits	Atts	Hit
1	14	word msd	vgrajen brezžični omrežni vmesnik Pdnmein Ppnmeid Pdnmeid Somei
2	12	word msd	izmerjeno primerjalno kontrastno razmerje Pdnsei Ppnsei Ppnsei Sosei
3	11	word msd	Ploska trinitronska Največja frekvenca Ppnzei Ppnzei Ppszei Sozei
4	11	word msd	Ploska navadna Največja frekvenca Ppnzei Ppnzei Ppszei Sozei
5	10	word msd	čistokrvna travna strelska akcija Ppnzei Ppnzei Ppnzei Sozei
6	9	word msd	zagotovljeno priljubljen večigralski način Pdnsei Ppnmein Ppnmeid Somei
7	8	word msd	zateženo postavljene nadaljevalne točke Pdnsei Pdnzmi Ppnmmt Sozer
8	8	word msd	pridodana zanimiva večigralska komponenta Ppnzei Ppnzei Ppnzei Sozei
9	8	word msd	dodatni dnevni fiksni strošek Ppnmeid Ppnmeid Ppnmeid Somei

Metadata searching

- Slovenian Biographical Lexicon
- Each article metadata about person
- Mounted under Fedora Commons

```
<listPerson>
  <person n="main">
    <sex value="1"/>
    <persName>
      <forename>Filip</forename>
      <surname>Abram</surname>
    </persName>
    <occupation>sodnik</occupation>
    <occupation>sodni upravnik</occupation>
    <birth>
      <date when="1835">1835</date>
      <date n="update" when="1835-04-19"/>
      <placeName n="reg">Štanjel</placeName>
    </birth>
    <death>
      <date when="1903-04-01">1. apr. 1903</date>
      <placeName n="reg">Dunaj = Wien [Avstrija]</placeName>
    </death>
  </person>
  <person n="author">
    <sex value="1"/>
    <persName key="Pc.">
      <forename>Janko</forename>
      <surname>Polec</surname>
    </persName>
  </person>
</listPerson>
```



1925

SLOVENSKI BIOGRAFSKI LEKSIKON

1991



Oseba: Spol/tip: * ?

Poklic:

Letnice: do rojstva / smrti

Kraj: †

Avtor:

Omembe:

Način: točno Vez: ALI



- [Predgovor](#)
- [Kolofon](#)

Seznam: [A](#) [B](#) [C](#) [Č](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [Š](#)
[T](#) [U](#) [V](#) [W](#) [Y](#) [Z](#) [Ž](#)

[Naključna oseba](#) [Seznam rodbin](#)
[Rojeni na današnji dan](#) [Umrli na današnji dan](#)

TEI and other meta-data standards

■ TEI SIG on Libraries

<http://wiki.tei-c.org/index.php/SIG:Libraries>

- ◆ Exploring TEI v.s. METS, MODS, EAD, MARC

■ TEI SIG on Ontologies

<http://wiki.tei-c.org/index.php/SIG:Ontologies>

- ◆ Exploring TEI v.s. CIDOC-CRM, FRBR
- ◆ Christian-Emil Ore and Øyvind Eide: TEI and cultural heritage ontologies: Exchange of information? *Literary and Linguistic Computing* 2009 24(2):161-17

Conclusions

- Presented metadata as encoded in the TEI header
- Some use cases & current work on mappings to other metadata standards
- TEI header has a rich structure – maybe too rich?
 - ◆ possible to encode same information in distinct ways
 - ◆ text content of TEI header elements can be further marked up