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

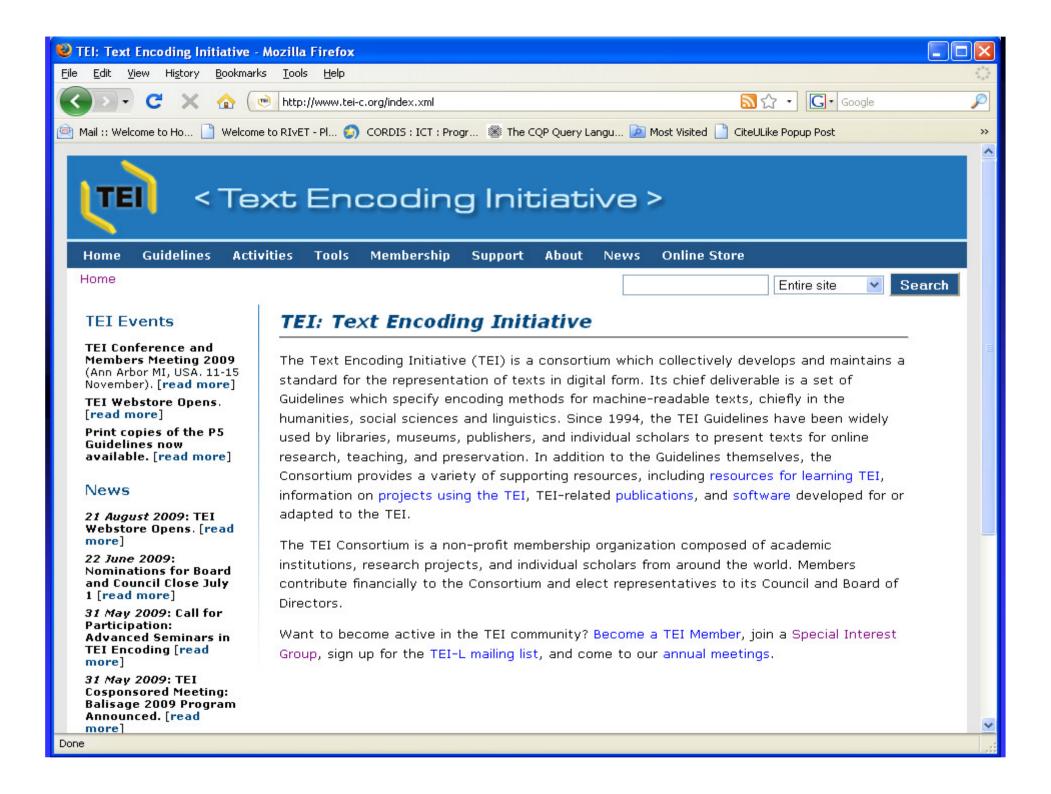
Perspectives on Metadata: Digital Editon & Preservation 12.11-13.11.2009, Vienna

Overview of the talk

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

Text Enoding 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 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>Znanstvenoraziskovalni center SAZU</publisher>
       <pubPlace>Ljubljana
       <date>2009</date>
       <idno type="ISBN">ISBN 978-961-268-001-5</idno>
       <availability> [4 lines]
       <pubPlace> [3 lines]
     <sourceDesc>
       <billStruct>
         <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 descritpions

- Škofja loka Passion Play: oldest performace 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">
                J11
                 <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 trascriptions 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>:

```
="21" id="bsDT.1.021">ba vúelica · Bofe gozpodi miloʒtiví · tebe ze mil</rr>
= "22" id="bsDT.1.022">tuori<g corresp="zrcolaEB81"/> · od · ʒih poftenih greh ·

= "23" id="bsDT.1.023">I · vuénfih · í minfih · Efe iezem ztvoril · teh ze · <g corr</p>
= "24" id="bsDT.1.024">miltuori<g corresp="zrcolaEB81"/> · I' · Zuetei marii · I'
```

TEI header <charDesc>:

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 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 xml:id="Ft.P.P.0">
    <catDesc>
    <term xml:lang="sl">objavljeno</term>
    <term xml:lang="en">published</term>
   </catDesc>
   <category xml:id="Ft.P.P.O.K">
     <catDesc>
    <term xml:lang="sl">knjižno</term>
    <term xml:lang="en">book</term>
   </catDesc>
```

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="#N0. #N1.c #N2.m #N3.s #N4.n"/>
<fs xml:id="Ncmsg" xml:lang="en" feats="#N0. #N1.c #N2.m #N3.s #N4.g"/>
<fs xml:id="Ncmsd" xml:lang="en" feats="#N0. #N1.c #N2.m #N3.s #N4.d"/>
<fs xml:id="Ncmsan" xml:lang="en" feats="#N0. #N1.c #N2.m #N3.s #N4.a #N5.n"/>
<fs xml:id="Ncmsay" xml:lang="en" feats="#N0. #N1.c #N2.m #N3.s #N4.a #N5.y"/>
```

TEI header feature library:

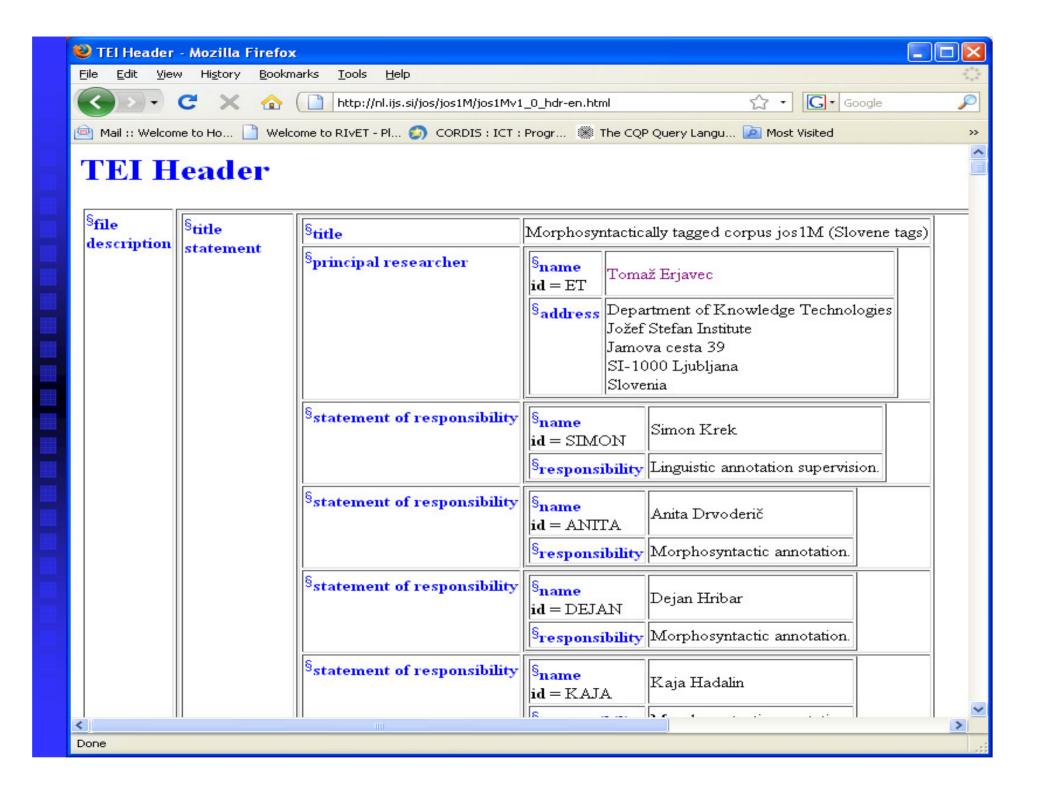
```
<fLib>
<!--1. Noun (N)-->
<f name="PoS" xml:id="N0." xml:lang="en"><symbol value="Noun"/></f>
<f name="Type" xml:id="N1.c" xml:lang="en"><symbol value="common"/></f>
<f name="Type" xml:id="N1.p" xml:lang="en"><symbol value="proper"/></f>
<f name="Gender" xml:id="N2.m" xml:lang="en"><symbol value="masculine"/></f>
<f name="Gender" xml:id="N2.f" xml:lang="en"><symbol value="feminine"/></f>
<f name="Gender" xml:id="N2.f" xml:lang="en"><symbol value="feminine"/></f>
<f name="Gender" xml:id="N2.n" xml:lang="en"><symbol value="feminine"/></f>
</fr>
```

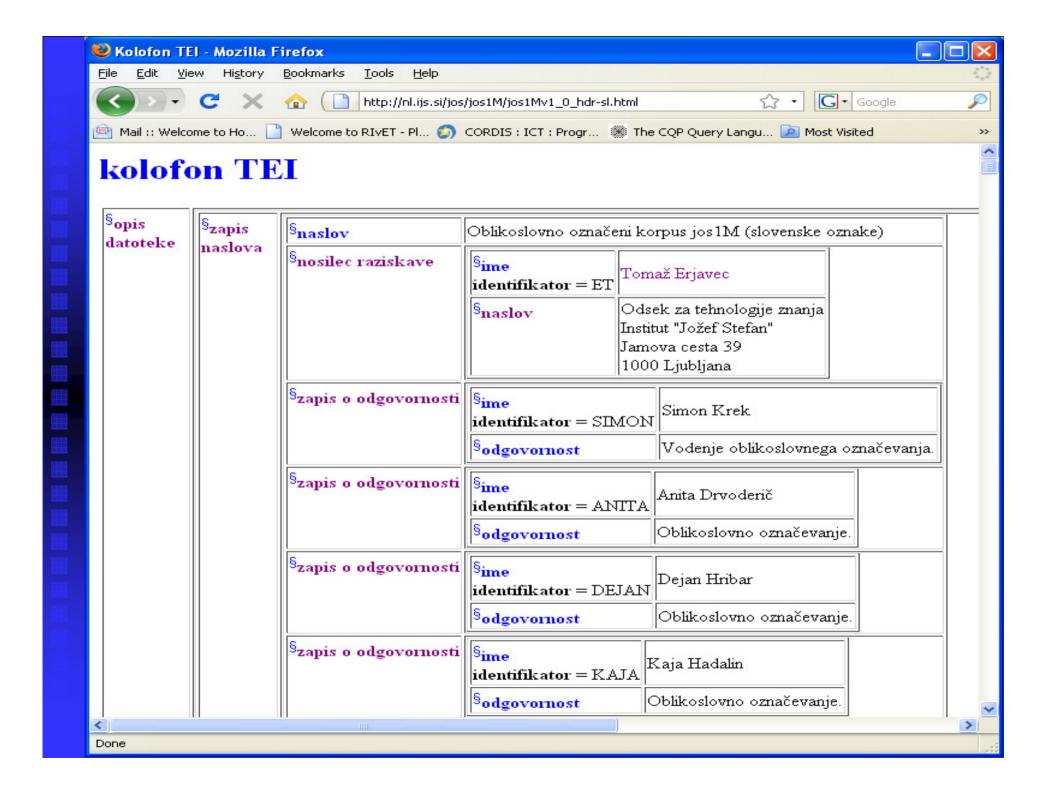
III. TEI metadata at work

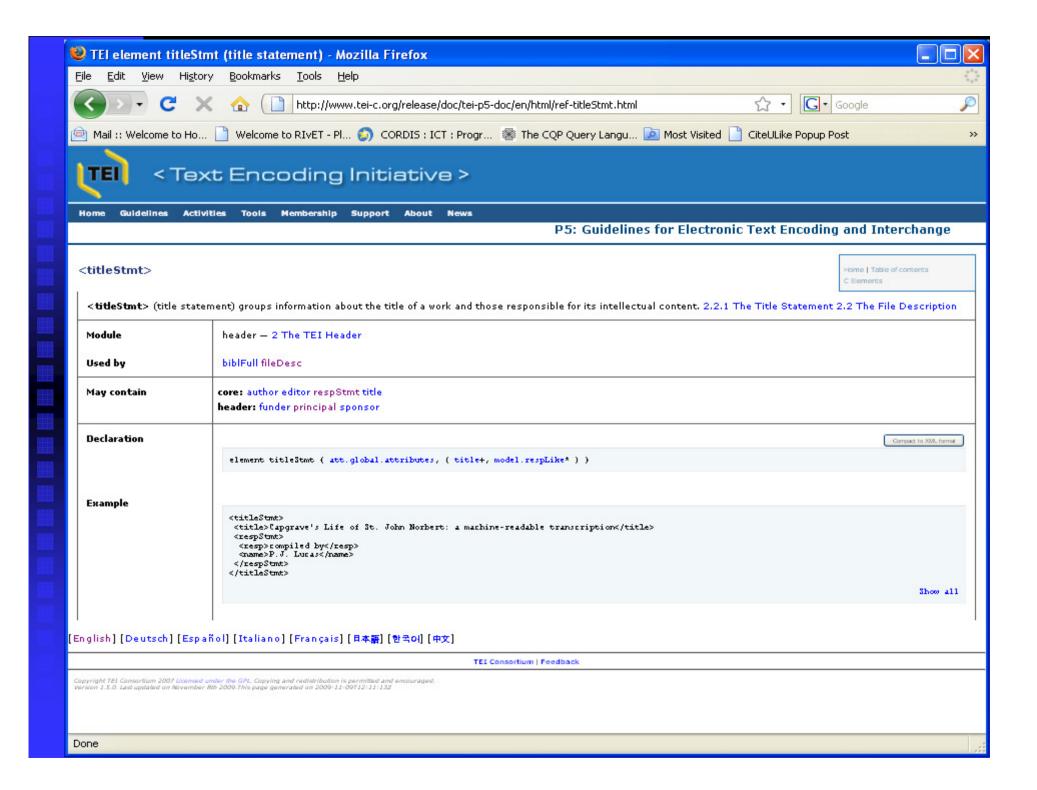
- 1. HTML rendering
- 2. Concordancer attributes
- 3. Repository search

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 langauge
- TEI elements connected to their definitions

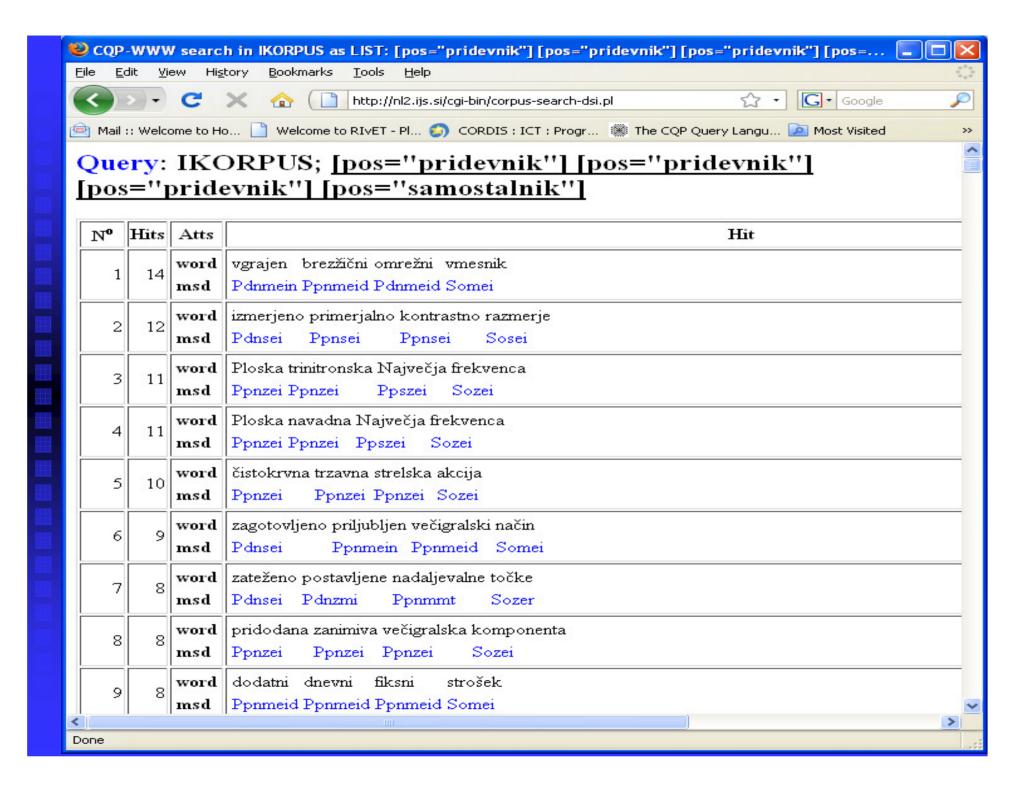






Concordance attributes

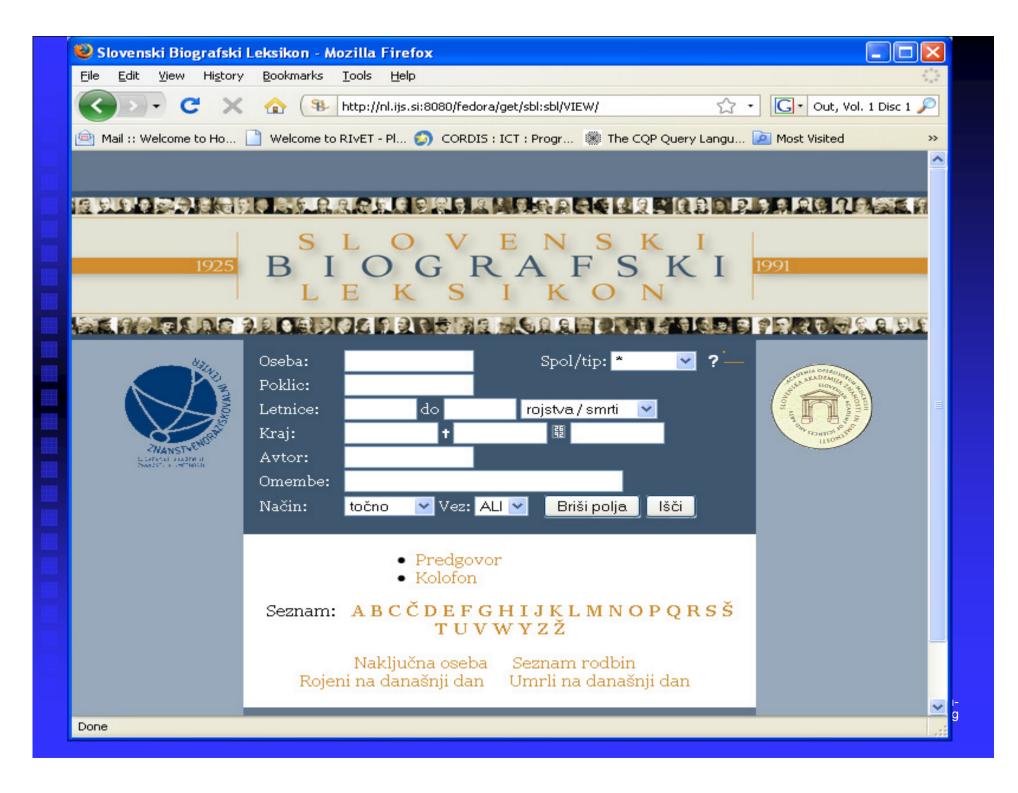
- Some concordancers can also search & display arbitrary positional attributes of tokens
- E.g. lemmea and tags such as "Ncms"
- With the feature library, it is simple to decompose such tags into features, e.g. PoS=Noun, Type=common, Gender=masculine, Number=singular
- This makes it possible to search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



Metadata searching

- SlovenianBiographicalLexicon
- Each article metadata about person
- Mounted under Fedora Commons

```
tPerson>
 <person n="main">
   <sex value="1"/>
   <persName>
      <forename>Filip</forename>
      <surname>Abram</surname>
   </persName>
    <occupation>sodnik</occupation>
    <occupation>sodni upravnik</occupation>
    <hirth>
     <date when="1835">1835</date>
     <date n="update" when="1835-04-19"/>
     <placeName n="reg">Štanjel</placeName>
   </hirth>
   <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 kev="Pc.">
      <forename>Janko</forename>
     <surname>Polec</surname>
   </persName>
 </person>
</listPerson>
```



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

Erjavec: Metadata as a Strategy of domainspecific Content Modelling

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