Standardizacija zapisa jeziko(slov)nih podatkov in TEI

Tomaž Erjavec

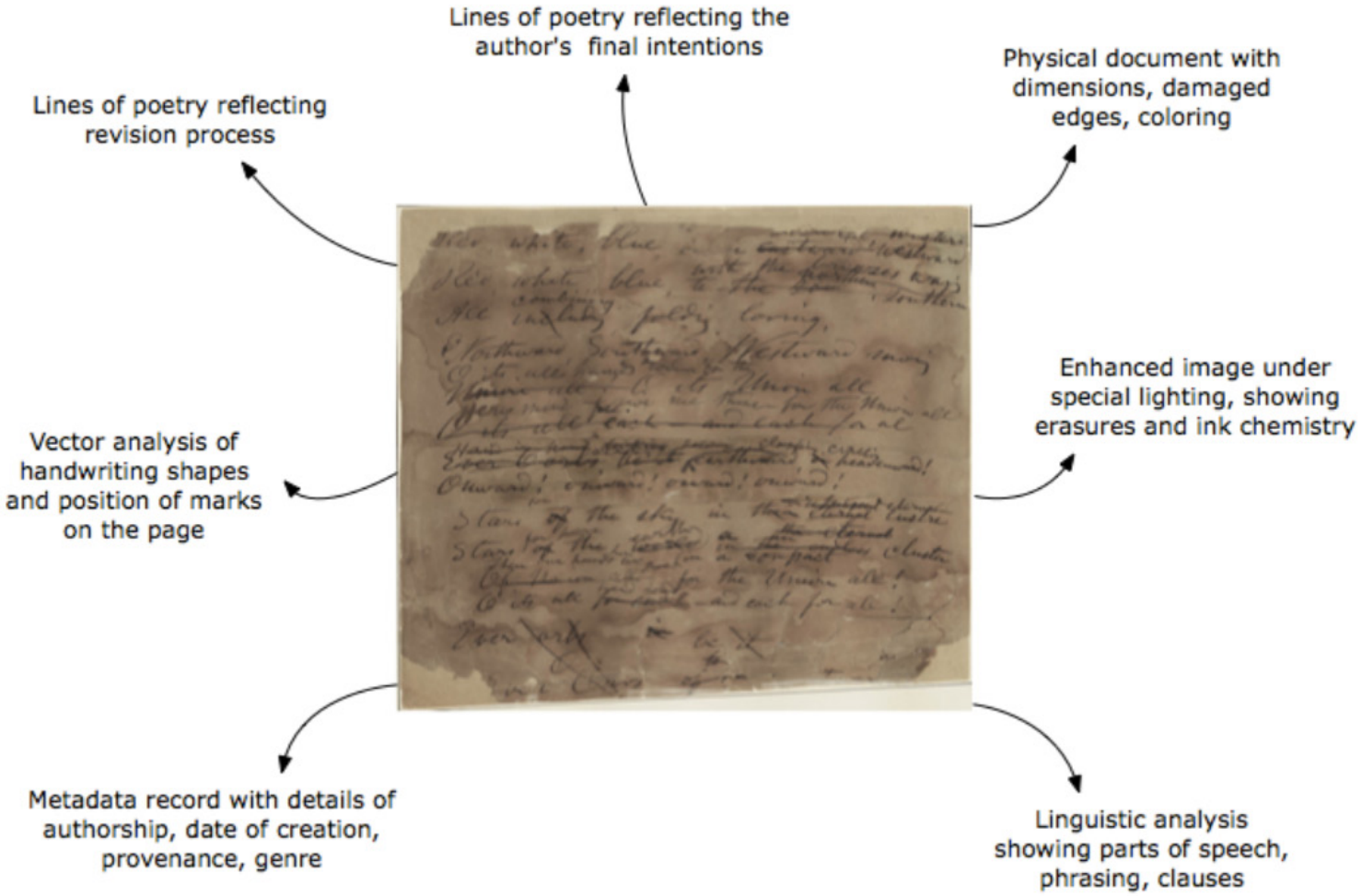
Odsek za tehnologije znanja, Institut „Jožef Stefan“

FRI, 10. 11. 2017

Pregled predavanja

- Uvod
- Zgodovina, organizacija in namen TEI
- Moduli TEI
- TEI v Sloveniji

Kodiranje besedila: predstavitev objektov preučevanja



https://www.wwp.northeastern.edu/outreach/seminars/context_2011-09/presentations/overview/overview_new_01.xhtml

Formati zapisov besedila

- slikovni formati
  (faksimile besedila)
- zvočni formati
  (govor)
- video
  (posnetek govorca)
- tabelarični formati
  (leksikoni, nekatere učne množice)
- **XML**
  (označena besedila)

Prednosti XML

- podpira mešanje besedila in oznak, hierarhične strukture, kazalci
- standard W3C
- formalna validacija prek sheme: DTD, W3C shema, RelaxNG, Schematron
- pridruženi standardi: XInclude, XPath, XSLT, XQuery, …
- in orodja: Saxon (XSLT), eXist (XQuery), urejevalniki (Oxygen, emacs, …)

- primeren predvsem kot arhivski format
- enostavna pretvorba v formate za aplikacije: HTML, JSON, tabele, PDF, …

XML sheme za označevanje besedil

- XML je meta-jezik: shema nato definira konkretne elemente, njihove atribute in dovoljena gnezdenja
- poleg formalne sheme (sintaksa) potrebujemo tudi razlago pomena elementov (semantika)
- shemo lahko naredimo za svoje namene sami
- za namene izmenjave podatkov (in podpore orodij) je bolje uporabiti eno od standardnih shem
- za označevanje jezikoslovnih kategorij obstaja veliko standardnih shem oz. dobrih praks:
  - ISO TC 37 SC4 standardi za kodiranje jezikoslovnih oznak
  - Nacionalni projekti (TCF: Nemčija, FoLiA: Nizozemska, …)
  - in TEI, vendar je ta mnogo širši

< Text Encoding Initiative >

- začet kot raziskovalni projekt humanistike
  - podprt s strani treh strokovnih društev
  - financiran 1990-1994 (ZDA, EU)
- vplivne zvrsti besedil
  - digitalne knjižnice, zbirke besedil
  - jezikovni korpusi
  - znanstvene podatkovne množice
- mednarodni konzorcij ustanovljen 1999
- spletna stran: http://www.tei-c.org/

Cilji TEI

- boljša izmenjava in integracija znanstvenih podatkov
- podpora vsem vrstam besedila, za vse jezike in obdobja
- napotki za začetnike: kaj kodirati: kodifikacija dobrih praks
- pomoč specialistom: kako kodirati - ohlapen okvir, v katerega lahko umestimo nepredvidljive razširitve

- Ti, na prvi pogled nekompatibilni cilji, so formalizirani v fleksibilnem in modularnem okolju.

Vodila TEI

- kaj „besedilo“ v resnici je
- kodifikacija sodobnih znanstvenih pristopov k analizi besedil
- konsenzualne predpostavke in prioritete o digitalni agendi:
  - poudarek na vsebini in funkciji (in ne predstavitvi)
  - identifikacija generičnih rešitev (in ne specifičnih za aplikacije)

Priporočila TEI

- priporočila za kodiranje besedil, ki pokrivajo tako generične strukture, kot tudi zelo specifična področja
- velika zbirka definicij XML elementov s pridruženimi deklaracijami za razne jezike shem XML
- v tisku ~1.200 strani
- modularen sistem za izdelavo namenskih shem XML
- priporočila TEI so napisana v TEI in združujejo dokumentacijo in formalne definicije (ODD: One Document Does it all)
- vzdržujejo se na GitHub, dostopna v TEI/XML, HTML, PDF, EPUB,…
- cf. http://www.tei-c.org/Guidelines/

Povezava z XML

TEI definira okvir za definicijo množice shem XML:

- definira imena in pomene nekaj sto koristnih besedilnih kategorij
- vključuje nabor modulov, ki jih uporabimo za definicijo shem, ki omogočajo te kategorizacije
- ponuja tudi mehanizem za prilagajanje in kombiniranje definicij z novimi, znotraj istega konceptualnega modela

Priporočila TEI P5 na spletu

Poglavja / moduli TEI

TEI ekosistem

- konzorcij TEI (člani, tehnični svet + svet direktorjev)
- zavezanost odprtemu dostopu in kolaborativnemu razvoju
- smernice in spletne strani redno vzdrževane
- prijazen in živahen dopisni seznam tei-l
- revija jTEI + redne letne konference TEI
- podporna orodja:
  - Roma: generiranje shem XML
  - TEI Stylesheets + OxGarage: pretvorba v in iz TEI

TEI v Sloveniji (1998-)

- projekti izgradnje referenčnih korpusov:
  Fida, FidaPLUS, Gigafida
- jezikovnotehnološki projekti EU (IJS et al.):
  MULTEXT-East, Concede, MondiLex; IMPACT
- projekti ZRC SAZU in SAZU + IJS:
  tekstnokritične izdaje slovenskega slovstva, slovenska biografija itd.
- projekti FF OAŠ + IJS:
  Japonsko-slovenski slovar in korpusi
- DARIAH-SI + CLARIN.SI:
  digitalna knjižnica + korpusi
- slovenski projekti:
  - JOS: jezikoslovno označevanje slovenščine
  - SSJ: sporazumevanje v slovenščini
  - Janes: jezikoslovna analiza nestandardne slovenščine
  - KAS: Slovenska znanstvena besedila: viri in opis

Primer uporabe TEI

G. Orwell: 1984

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader xml:lang="en" xml:id="mteo-sl.teiHeader"> [340 lines]
  <text xml:id="mteo-sl." xml:lang="sl">
    <body xml:id="Osl">
      <div xml:id="Osl.1" n="1" type="part">
        <head>Prvi del</head>
        <div xml:id="Osl.1.2" n="1" type="chapter">
          <head>I</head>
          <p xml:id="Osl.1.2.2">
            <s xml:id="Osl.1.2.2.1">Bil je jasen, mrzel aprilski dan in ure so bile trinajst.</s>
            <s xml:id="Osl.1.2.2.2"><name>Winston Smith</name> je imel brado zakopano v prsi, da bi
              ušel strupenemu vetru, ko je stopil skozi steklena vrata bloka Zmaga, vendar ne dovolj
              hitro, da ne bi vrtinec peščenega prahu vstopil skupaj z njim.</s>
          </p>
          <p xml:id="Osl.1.2.3">
            <s xml:id="Osl.1.2.3.1">Veža je smrdela po kuhanem zelju in starih, cunjastih
              predpražnikih.</s>
            <s xml:id="Osl.1.2.3.2">Na eni strani je bil na steno pribit barven, za notranjo opremo
              prevelik plakat.</s>
            <s xml:id="Osl.1.2.3.3">Prikazoval je preprosto ogromen, več kot meter velik obraz:
              obraz moškega pri petinštiridesetih, s košatimi črnimi brki in z ostro začrtanimi,
              čednimi potezami.</s>
```

I. Cankar: S poti

I. [BENETKE]{VENEZIA}.

§1 Težko si je misliti večjo razliko, nego je med ljubljanskim septembrskim jutrom in benečanskim septembrskim večerom. Zrak se mi vtihotaplja v sobo kakor lepa godba v srce, kadar sem obupal nad samim seboj in se povprašujem, čemu sem na svetu. In vendar je tudi današnje ljubljansko jutro bilo lepo, kljub tisti mrzli mokroti, ki peče oči. [Snoči sem bil pri izpovedi; prvo prebujenje po izpovedi je nekaj neizrečeno prijetnega. Človek odpre oči z zavestjo otroške dušne svežosti; sam s seboj je nežen kakor mati; preden prižge svečo, je tema vsa prepredena s svetlobo; svet je tako skladen in dober, vsakdo posebej poljuba vreden; kar je bilo srcu veliko, mu je majhno in le eno potrebno: ne delati krivice. Vrhutega so mi bili računi že v redu, obresti poravnane, drobne skrbi odložene. Ko sem se vozil na kolodvor, mi je cigareta neizmerno teknila.]

§2 Na peronu {me} je že čakal moj mladi prijatelj Fritz, poet in umetnostni zgodovinar iz rajha. Prezebal je kljub svoji topli športni suknji. Pogled mu je bil v kolodvorskih meglah, parah in sajah kakor vlažen protest. Bilo je gotovo, da je slabo spal.

§3 „Ne, sploh nisem spal. Vaši hoteli so brlogi za najbolj pokore potrebne [izpokornike]{spokornike}. Večerjal sem žlico juhe, v postelji sem si zlomil vrat in dobil revmatizem. Bagage.“

```
<p id="p.3" n="3">„Ne, sploh nisem spal. Vaši hoteli so brlogi za najbolj
    pokore potrebne <app>
        <rdg wit="DS">izpokornike</rdg>
        <rdg wit="KN">spokornike</rdg>
    </app>. Večerjal sem žlico juhe, v postelji sem si zlomil vrat in dobil
    revmatizem. Bagage.“</p>
```

Slovenska biografija: splet

Slovenska biografija: TEI

```
<person xmlns="http://www.tei-c.org/ns/1.0" xml:id="sbi619828"
    corresp="sbl-text.xml#sbl03314" role="main">
    <idno type="URL">http://www.slovenska-biografija.si/oseba/sbi619828/</idno>
    <sex value="1"/>
    <persName>
        <forename>Jan</forename>
        <surname>Strgar</surname>
    </persName>
    <persName>
        <forename>Janez</forename>
        <surname>Strgar</surname>
    </persName>
    <persName>
        <forename>Jan</forename>
        <surname>Stergar</surname>
    </persName>
    <occupation scheme="#occupation" code="#cebelar"/>
    <occupation scheme="#occupation" code="#trgovec"/>
    <birth>
        <date when="1881-05-08">8. maja 1881</date>
        <placeName>
            <settlement>Nemški Rovt</settlement>
            <country>Slovenija</country>
            <geo>46.2713028 13.9787442</geo>
        </placeName>
    </birth>
```

Seje slovenskega parlamenta 1990-1992

```
<person xml:id="ZupancicJoze1936">
   <persName>
      <surname>Zupančič</surname>
      <forename>Jože</forename>
   </persName>
   <sex value="M"/>
   <birth when="1936"/>
   <occupation when="1990-04">prof. geografije</occupation>
   <affiliation when="1990-04">
      <orgName>Gimnazija Celje</orgName>
   </affiliation>
   <residence when="1990-04">Kopitarjeva ulica 2, Celje</residence>
   <affiliation ref="#parl.ZbZdruDel" ana="#parl.Skup-11 #grp.member">
      <state>
         <label>izobraževanje in telesna kultura</label>
      </state>
   </affiliation>
   <affiliation ref="#independet" ana="#parl.Skup-11"/>
</person>
```

```
<u xml:id="ZbZdruDel.1990-05-07.s001-01.sp-34.3" who="#ZupancicJoze1936"
   ana="#taxonomy.root #topic.1">Sedaj pa drugo. Mislim, da bi lahko v skladu s 33.
   členom tega poslovnika tudi nekako ugotovili, lepo vas prosim, ne bi rad, da me
   napačno razumete, da bi vseeno bilo nekoliko bolj precizno urejeno vprašanje kluba
   neodvisnih poslancev za vse poslance, ki niso strankarsko vezani in nimajo matične
   stranke, ki bi skrbele za njihovo strokovno in še kakšno drugo dejavnost. Nekaj nas
   bo takšnih, ki smo samohodci, samostrelci, pa še mogoče kaj drugega. Če smo samo
```

jaSlo japonsko-slovenski slovar

bakeru ばける【化ける】(V1)

pojaviti se pod krinko; *preleviti se*; *prevzeti obliko*; *spremeniti se (v pošast)*

- 狐（きつね）が女の子（おんなのこ）に化けた。
  Lisica se je spremenila v deklico.
- 化け猫（ばけねこ）
  (čarobna) ***mačka****, ki se je spremenila [v človeka ipd.]*

težavnostna stopnja 1

konkordance za ばける: L2 (3), L0 (9), jpWaC (174)

konkordance za 化ける: L1 (24), L0 (108), jpWaC (743)

```
<entry xml:id="jaslo.8547">
   <form xml:lang="ja" type="hw">
      <orth type="roma">bakeru</orth>
      <orth type="kana">ばける</orth>
      <orth type="kanji">化ける</orth>
   </form>
   <gramGrp>
      <pos>V1</pos>
   </gramGrp>
   <cit type="translation" xml:lang="sl">
      <quote xml:lang="sl">pojaviti se pod krinko</quote>
```

ssj500k v2.0: učni korpus

```
<p xml:id="ssj1.1">
   <s xml:id="ssj1.1.1">
      <pc xml:id="ssj1.1.1.t1" ana="msd:U">"</pc>
      <w xml:id="ssj1.1.1.t2" ana="msd:Zk-mer" lemma="tisti">Tistega</w>
      <c> </c>
      <w xml:id="ssj1.1.1.t3" ana="msd:Somer" lemma="večer">večera</w>
      <c> </c>
      <w xml:id="ssj1.1.1.t4" ana="msd:Gp-spe-n" lemma="biti">sem</w>
      <c> </c>
      <w xml:id="ssj1.1.1.t5" ana="msd:Rsn" lemma="preveč">preveč</w>
      <c> </c>
      <w xml:id="ssj1.1.1.t6" ana="msd:Ggdd-em" lemma="popiti">popil</w>
      <pc xml:id="ssj1.1.1.t7" ana="msd:U">,</pc>
      <c> </c>
      <phr type="IReflV">
         <w xml:id="ssj1.1.1.t8" ana="msd:Ggdd-es" lemma="zgoditi">zgodilo</w>
         <c> </c>
         <w xml:id="ssj1.1.1.t9" ana="msd:Zp------k" lemma="se">se</w>
      </phr>
      <c> </c>
      <w xml:id="ssj1.1.1.t10" ana="msd:Gp-ste-n" lemma="biti">je</w>
```

ssj500k v2.0: skladnja

```
<linkGrp type="syntax" targFunc="head argument" corresp="#ssj1.1.1">
   <link ana="syn:modra" target="#ssj1.1.1 #ssj1.1.1.t1"/>
   <link ana="syn:dol" target="#ssj1.1.1.t3 #ssj1.1.1.t2"/>
   <link ana="syn:štiri" target="#ssj1.1.1.t6 #ssj1.1.1.t3"/>
   <link ana="syn:del" target="#ssj1.1.1.t6 #ssj1.1.1.t4"/>
   <link ana="syn:tri" target="#ssj1.1.1.t6 #ssj1.1.1.t5"/>
   <link ana="syn:modra" target="#ssj1.1.1 #ssj1.1.1.t6"/>
   <link ana="syn:modra" target="#ssj1.1.1 #ssj1.1.1.t7"/>
   <link ana="syn:modra" target="#ssj1.1.1 #ssj1.1.1.t8"/>
   <link ana="syn:del" target="#ssj1.1.1.t8 #ssj1.1.1.t9"/>
   <link ana="syn:del" target="#ssj1.1.1.t8 #ssj1.1.1.t10"/>
   <link ana="syn:štiri" target="#ssj1.1.1.t8 #ssj1.1.1.t11"/>
```

goo300k: stara slovenščina

```
<choice>
   <orig>
      <w>ludy</w>
   </orig>
   <reg>
      <w lemma="človek" ana="#Ncm">ljudi</w>
   </reg>
</choice>
<c> </c>
<lb/>
<choice>
   <orig>
      <w>memujete</w>
   </orig>
   <reg>
      <w lemma="mimo" ana="#Rgp">mimo</w>
      <c> </c>
      <w lemma="iti" ana="#Vmb">iti</w>
   </reg>
</choice>
<pc>,</pc>
<c> </c>
<w lemma="biti" ana="#Va">je</w>
<c> </c>
```

Zaključki

- TEI v resnici ni drugega kot bogat nabor elementov in atributov za kodiranje raznovrstnih besedil in načinov njihove analize
- dokumentiran, preverljiv, opremljen z orodji in veliko skupnostjo uporabnikov
- vendar kompleksen in neperskriptiven (v škodo semantični interoperabilnosti)
- v svetu najbolj prodrl pri kompleksnih digitalnih izdajah
- v Sloveniji pa tudi v korpusnem in računalniškem jezikoslovju