RIDICS

Standard XML TEI P5 a veršované texty

Boris Lehečka, boris@daliboris.cz

Příspěvek byl podpořen projektem Ministerstva školství, mládeže a tělovýchovy č. LM2015081 „Výzkumná infrastruktura pro diachronní bohemistiku“ (akronym RIDICS, http://vokabular.ujc.cas.cz).

Ústav pro jazyk český
Akademie věd České republiky

FAKULTA ELEKTROTECHNICKÁ ČVUT V PRAZE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

XML

- e**X**tensible **M**arkup **L**anguage (rozšiřitelný značkovací jazyk)
- od roku 1998
  - verze 1.0, páté vydání (2008)
  - verze 1.1, druhé vydání (2006)
- původně pro textové dokumenty (nakladatelství)
  - využívá se pro výměnu dat

XML – značkování

- označení části textu, která obsahuje důležitý údaj (význam, nikoli forma)
- element (značka, tag)
  - určuje začátek a konec důležité části
  - `<title>Titul</title>`
  - `<pb />` (prázdný element; označuje hranici stránky)
- atribut
  - přidává ke značce doplňující údaje
  - `<title rend="center">Titul</title>`; `<pb n="[I]" />`

Pravidla pro značky a atributy

- názvy elementů a atributů nesmějí obsahovat mezeru
- atribut může být v rámci elementu pouze jednou
- každý element musí mít počáteční a koncovou značku
- prázdný element: <pb /> = <pb></pb>
- značky se nesmějí překrývat
  - správně: <hi rend="spaced">Konec.</hi>
  - špatně: <hi rend="spaced">Konec.** </hi>**

XML – jmenný prostor

- definuje, k jaké sadě značek elementy, popř. atributy patří
- zanořené elementy přebírají jmenný prostor nadřízeného elementu
- v jednom dokumentu lze kombinovat prvky z různých oblastí značkování
  - pomocí jmenných prostorů
- výchozí jmenný prostor
- další jmenné prostory s přiřazenou předponou

XML – jmenný prostor (ukázka)

```
<person
xmlns="http://www.tei-c.org/ns/1.0">
  <persName>
    <forename>Boris</forename>
  </persName>
</person>
```

```
<person
xmlns="http://docbook.org/ns/docbook">
  <personname>
    <firstname>Boris</firstname>
  </personname>
</person>
```

XML – jmenný prostor (předpona)

```
<person xmlns="http://www.tei-c.org/ns/1.0">
  <persName>
    <forename>Boris</forename>
   </persName>
</person>
```

```
<tei:person xmlns:tei="http://www.tei-c.org/ns/1.0">
  <tei:persName>
    <tei:forename>Boris</tei:forename>
  </tei:persName>
</tei:person>
```

XML a databáze

- podobnost s databází
  - význam = název pole (např. signatura, instituce, místo vydání)
- rozdíly
  - databáze: pevná struktura (vše je označeno)
  - XML: volná struktura (kombinace označených a neoznačených částí)
  - databázi lze převést na XML, naopak je to obtížnější až nemožné

XML – co a jak označovat

- repertoár značek a atributů
  - uzuální použití (na základě dohody)
- formální definice
  - pomocí DTD, XSD, Relax NG
  - pro kontrolu dokumentu
  - pro počítačové nástroje
    - nabízejí přípustné elementy, hodnoty atributů, kontrolují správnost

XML – software

- textové editory (s doplňky)
  - Notepad++ – https://notepad-plus-plus.org
  - jEdit – www.jedit.org
- specializované editory
  - XML Copy Editor – https://sourceforge.net/projects/xml-copy-editor/
  - Oxygen XML Editor – www.oxygenxml.com
  - Transkribus – https://transkribus.eu

XML TEI

- TEI = Text Encoding Initiative (základy v roce 1987)
- „spravuje standard pro reprezentaci textů v digitální podobě“
- první doporučení v roce 1993
  - *Guidelines for the Encoding and Interchange of Machine-readable Texts*, C.M. Sperberg-McQueen and L. Burnard, eds. (Chicago and Oxfrod, ACH-ACL-ALLC Text Encoding Initiative, 1993)
- TEI P5, verze 3.4.0 (23. 7. 2018), rev. 1fa0b54 (1. verze P5 v roce 2007)

XML TEI P5 – nápověda

- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
- společné elementy
  - odstavce, zvýraznění, citace, seznamy, odkazování, grafika, bibliografie
- specifické elementy
  - verše; divadelní hry; slovníky; rukopisy; textově-kritický aparát; jména, data, lidé, místa; tabulky, rovnice, noty; jazykový korpus…
- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html

XML TEI – základy

- popis dokumentu pomocí metadat (<teiHeader>)
- přepis textu dokumentu (<text>)
  - začátek (<front>): titulní strana, abstrakt, předmluva, dedikace…
  - hlavní text (<body>)
  - závěr (<back>): závěrečné pasáže (věnování, rejstříky, reklama)

XML TEI – hlavička

- popis dokumentu pomocí metadat (<teiHeader>)
  - autor, název, tiskař, datace
  - uložení tisku
  - fyzický popis artefaktu (rozměry, materiál, grafická výzdoba)
  - informace o přepisu (editor, použité transkripční zásady)
  - digitální kopie (obrázky)
- viz Špalíček, záložka XML

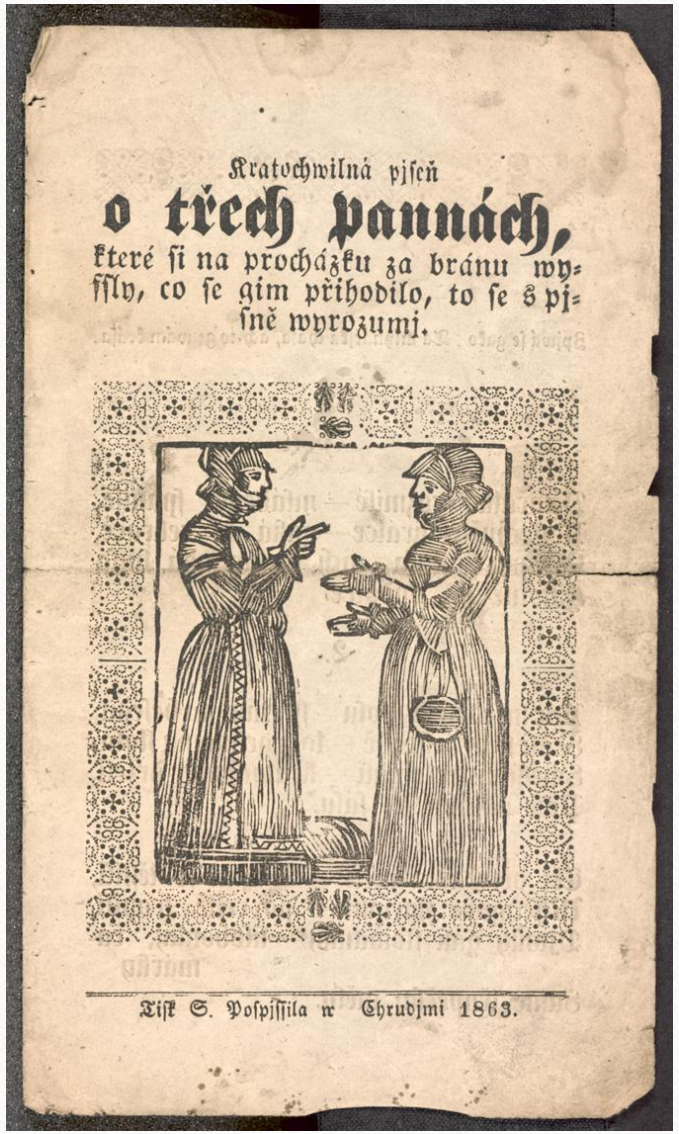
XML TEI – přepis textu

- zachytit ideální výsledek (bez chyb)
  - + upozornit na chybná místa

```
<choice>
   <corr>ſſprýmy</corr>
   <sic>ſſprými</sic>
</choice>
```

- zachytit věrně originál
  - + uvést opravené znění u chyb

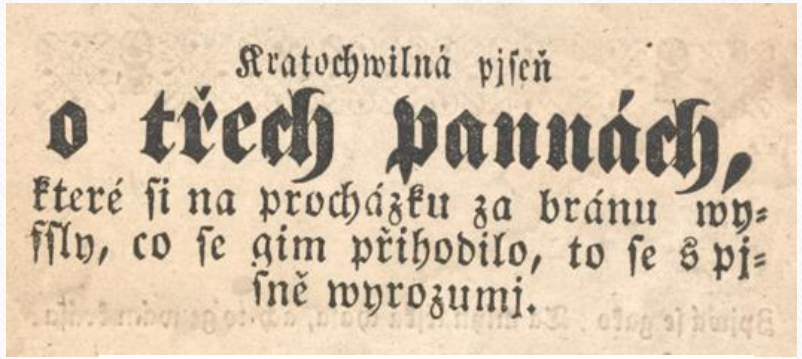
```
<choice>
  <orig>ſſprýmy</orig>
  <reg>ſſprými</reg>
</choice>
```




```
<front>
  <pb n="[1]" facs="#KP_D_39_1_____0CIIN30001P"/>
  <titlePage rend="center">
  <titlePart>Kratochwilná pjſeň <lb />
      <hi rend="big">o třech pannách,</hi> <lb />
      které ſi na procházku za bránu wy=<lb break="no"  />
      ſſly, co se gim přihodilo, to ſe s pj=<lb break="no"  />
      ſně wyrozumj.
   </titlePart>
   <figure />
   <docImprint rend="center">Tiſk <publisher>S.
Poſpjſſila</publisher> w <pubPlace>Chrudjmi</pubPlace>
<date>1863</date>.</docImprint>
  </titlePage>
 </front>
```

- <pb /> = (page beginning); před jakýmkoli textem/elementem
- @n = číslo strany (v [] = doplněno editorem)
- @facs = odkaz na obrázek strany

Titul


Kratochwilná pjſeň
o třech pannách,
které ſi na procházku za bránu wy=
ſſly, co ſe gim přihodilo, to ſe s pj=
ſně wyrozumj.

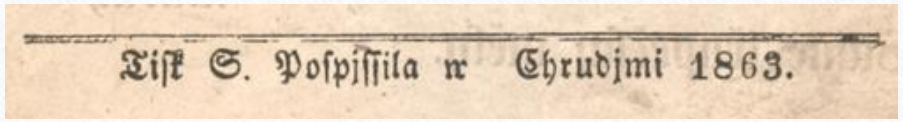
Kratochwilná pjſeň
▹o třech pannách,◃
které ſi na procházku za bránu wy=
ſſly, co se gim přihodilo, to ſe s pj=
ſně wyrozumj.
figure

Tiſk ▹S. Poſpjſſila◃ w ▹Chrudjmi◃ ▹1863◃.

```
<titlePart>Kratochwilná pjſeň <lb />
    <hi rend="big">o třech pannách,</hi> <lb />
    které ſi na procházku za bránu wy=<lb break="no" />
    ſſly, co se gim přihodilo, to ſe s pj=<lb break="no" />
    ſně wyrozumj.
</titlePart>
```

- <lb /> = začátek řádku (line beginning)
- <hi> = zvýraznění (highlighted)
- pomocí @rend lze definovat způsob zvýraznění

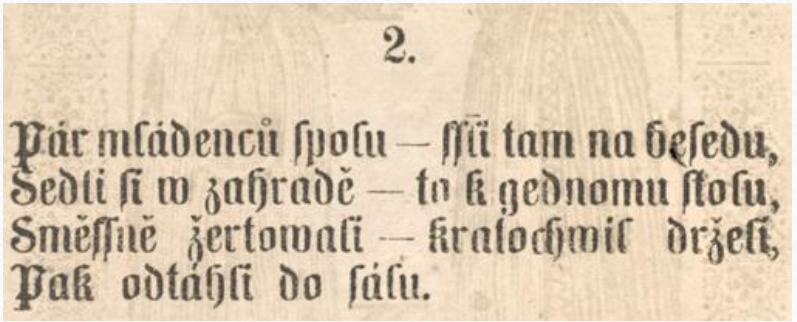
Tiráž


Tiſk S. Poſpjſſila w Chrudjmi 1863.

```
<docImprint rend="center">Tiſk
 <publisher>S. Poſpjſſila</publisher>
 w
 <pubPlace>Chrudjmi</pubPlace>
<date>1863</date>
.</docImprint>
```

- podchycené údaje o tiskaři, místu a datu vydání

Sloky, verše



2.

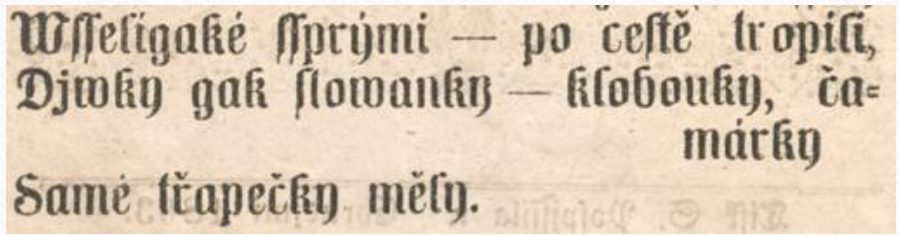
Pár mládenců ſpolu — ſſli tam na beſedu,
Sedli ſi w zahradě — to k gednomu ſtolu,
Směſſně žertowali — kratochwil drželi,
Pak odtáhli do ſálu.

- <lg> = skupina veršů (line group); pomocí @type lze upřesnit typ (např. refrén)
- <head> = nadpis

```
<lg type="stanza">
<head rend="center">2.</head>
<l n="5">Pár mládenců ſpolu – ſſli tam na beſedu,</l>
<l n="6">Sedli ſi w zahradě – to k gednomu ſtolu,</l>
<l n="7">Směſſně žertowali – kratochwjl drželi,</l>
<l n="8">Pak odtáhli do ſálu.</l>
</lg>
```

- <l> = verš (verse line); @n označuje číslo verše (jeho pořadí)

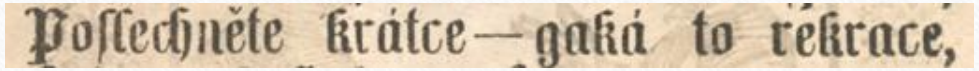
Opravy, doplnění


Wſſeligaké ſſprými — po ceſtě tropili,
Djwky gak ſlowanky — klobouky, ča-
márky
Samé třapečky měly.

- <corr> = opravený text
- <sic> = původní, chybný text
- <supplied> = text doplněný editorem;
- pomocí @resp lze určit editora, který zásah provedl

```
<l n="11">Djwky gak <choice>
 <corr>Slowanky</corr>
  <sic>ſlowanky</sic>
 </choice> – klobouky,
čamárky<supplied
resp="#BL">,</supplied></l>
```

Poznámka



- <note> = poznámka;
- pomocí @n lze zadat číslo, pod nímž se objeví ve výstupu;
- pomocí @resp lze určit editora, který poznámku vložil

```
<l n="2">
Poflechněte krátce – gaká to rekrace
<note n="1" resp="#BL">tj.
legrace</note>
,</l>
```

Další důležité prvky

- = oddíl; slouží k seskupení souvisejících prvků
- <add> = vložený text (autorem, korektorem, nikoli editorem)
- <c> = litera, např. iniciála
- <cb /> = hranice sloupce
- <del> = odstraněný text
- <expan> = rozepsání zkratky
- <fw> = prvky na stránce mimo hlavní text (záhlaví, zápatí, okraje)
- <foreign> = cizojazyčný prvek v textu
- <milestone /> = hranice v textu, kterou nelze zachytit strukturním elementem
- <quote> = citace (odkazující mimo text)
- <rhyme> = rýmující se výraz ve veršovaném textu
- <subst> = nahrazení jednoho textu jiným
- <unclear> = nečitelné místo v prameni

Děkuju za pozornost

- e-mail: boris@daliboris.cz
- Vokabulář webový: http://vokabular.ujc.cas.cz