Iniciatíva kódovania textu (TEI)

Prof. PhDr. Dušan Katuščák, PhD.

Slezská univerzita, FPF-ÚBK Opava

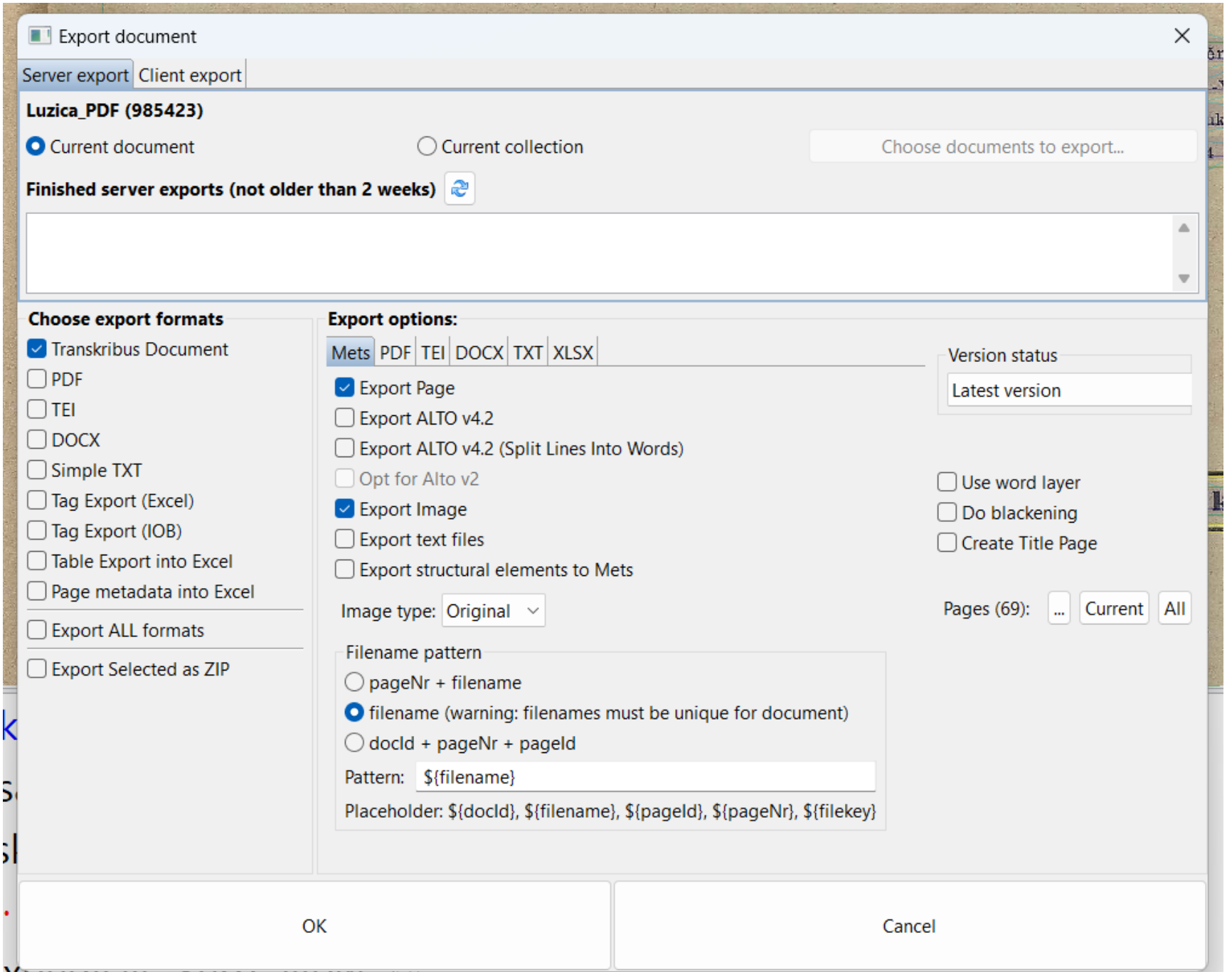
1-1 Wukhadźa· srjedź· kóždeho↵

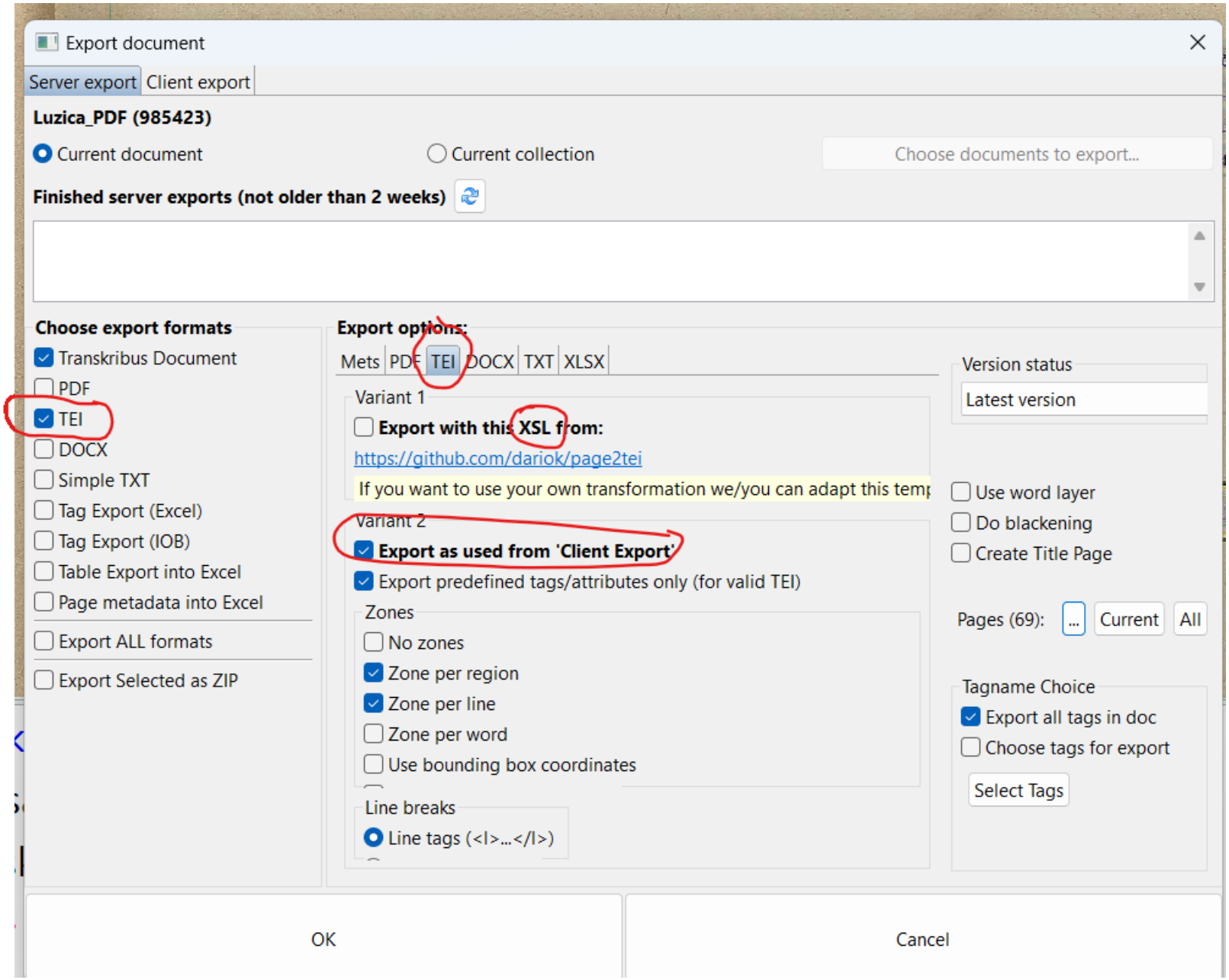
1-2 měsaca· w· Budyšinje· a· hodźi↵

1-3 so· skazać· pak· na· póstach,↵

1-4 pak· pola· administratora· a↵

1-5 w· expediciji· „Serb.· Now.“↵


Export document
Server export
Client export
Luzica_PDF (985423)
Current document
Current collection
Choose documents to export...
Finished server exports (not older than 2 weeks)
Choose export formats
Transkribus Document
PDF
TEI
DOCX
Simple TXT
Tag Export (Excel)
Tag Export (IOB)
Table Export into Excel
Page metadata into Excel
Export ALL formats
Export Selected as ZIP
Export options:
Mets
PDF
TEI
DOCX
TXT
XLSX
Export Page
Export ALTO v4.2
Export ALTO v4.2 (Split Lines Into Words)
Opt for Alto v2
Export Image
Export text files
Export structural elements to Mets
Image type:
Original
Version status
Latest version
Use word layer
Do blackening
Create Title Page
Pages (69):
...
Current
All
Filename pattern
pageNr + filename
filename (warning: filenames must be unique for document)
docId + pageNr + pageId
Pattern:
${filename}
Placeholder: ${docId}, ${filename}, ${pageId}, ${pageNr}, ${filekey}
OK
Cancel


Export document
Server export
Client export
Luzica_PDF (985423)
Current document
Current collection
Choose documents to export...
Finished server exports (not older than 2 weeks)
Choose export formats
Transkribus Document
PDF
TEI
DOCX
Simple TXT
Tag Export (Excel)
Tag Export (IOB)
Table Export into Excel
Page metadata into Excel
Export ALL formats
Export Selected as ZIP
Export options:
Mets
PDF
TEI
DOCX
TXT
XLSX
Variant 1
Export with this XSL from:
https://github.com/dariok/page2tei
If you want to use your own transformation we/you can adapt this temp
Variant 2
Export as used from 'Client Export'
Export predefined tags/attributes only (for valid TEI)
Zones
No zones
Zone per region
Zone per line
Zone per word
Use bounding box coordinates
Line breaks
Line tags (<l>...</l>)
Version status
Latest version
Use word layer
Do blackening
Create Title Page
Pages (69):
...
Current
All
Tagname Choice
Export all tags in doc
Choose tags for export
Select Tags
OK
Cancel

1 z 10 466

Export document complete

Doručené ×

transkribus@uibk.ac.at 22:08 (pred 0 minútami)

komu: mne

Dear Dusan,
your exported documents can be downloaded at:

https://transkribus.eu/export/3013536798539020918/export_job_4598999.zip
(this link expires in two weeks)

Exported documents (collection 183568):

Luzica_PDF (id=985423)

Best wishes

The Transkribus team

Received, thank you. | Thank you! | Thanks a lot.

Wukhadža srjedź kóždeho měsaca w Budyšinje a hodźi so skazać pak na póstach, pak pola administratora a w expediciji „Serb. Now.“

ŁUŽICA.

Płaći lětnje w Němskej: 3 hriwny, w Rakusko-Wuherskej: 4 krony, we Wukhodže: 2 rublej, w Zapadže: 4 franki.

Měsačnik za zabawu a powučenje.

Zhromadny časopis hornjo- a delnjołužiskich Serbow.

Z pomocu knjeza kanonika **Jakuba Herrmanna** we Wotrowje
wudawa **Jakub Bart-Ćišinski** w Pančicach.
Administrator: wučeŕ **Mikławš Hajna** w Konjecach.

Čisło 1. **Januar 1909.** **Lětnik 28.**

W pruzy njebjeskej.

Cyklus basni Jakuba Ćišinskeho.

Lubinska sonina.

Kaž we snje Lubin stoji z wječora —
A njebjo zamróčene woko ma —
Ze čŕjódami kaž k smjerći wokoł' hory
A z na trach krakanjom so honja wrony:
A delka płakaju na prózdne dwory
A noša styšć přez štomow stare krony.

Po polach stona wokoł' hory delka,
Hdźež lóšt bě był a było lěhwo wjelka.
A z hłubin na wjeŕch zazdychuja styski,
Hdźež ržale běchu modlitwy a wyski ...

Jow serbskich ludow zabitych je row
A žałosćacych wutrobow nět skhow. —

Přez jědle, šmrěki wichor wrjeskota,
Kaž bychu wojownicy přez lěs hnali;
A wjeŕski koći, k zemi praskota,
Kaž z mječom do škitow sej bychu prali ...

A zdobom kaž by wichor wotdychnył,
'noż ćeńki płač by słyšeć z rowow był. —

A swětło ze skały na wjeŕchu prasny
A wobraz w nim so zjewi čisty, jasny:
Tam žónska stoji stawow hoberskich,
Ta poslednja ze žonow wěšćeŕskich.

Płašć wot ramjenjow po postawje pada
Kaž z moška mjehki tak a běły tkany;
A z włosow błyšć kaž z pruhow słónčnych rjany
Ju z koškom spěšnym majka po khribjeće;
A króna dejmanty do złota pleće.
A woheń z kuzłaŕskeju woćow hlada,
Kiž sylzy z njeju do rubinow twori;
A ze rta jako płomjo ji so hori.
Nad hłowu běły sokoł křidle šěri

A z wótrym wokom do dala so měri;
A z hłowy stawa jemu pjerjow błyšć,
Zo dótknyć njemohł so ni strach ni styšć.

A žónska měša w kótle dźiwny war,
Pak khwalba ze rta šepce ji pak swar.
A z hłosom k njebjesam so wola wona
A z hóŕkim płačom čiši styskne hrona.
A wokoł' kótła z dźiwjej nohu khodźi,
Prut złoty do kótła, pak z kótła wodźi.
A jako sokoł woko wupina
Do štyrjoch stronow dele z Lubina ...

A hišće z prutom złotym w kótle měša,
Zo jako krew so horca na njón wěša.
A krjepi wysoko přez wšitke štomy
Na serbske dele wutroby a domy
Jich starych bohow móc a měr a škit,
Kaž činić póslał bě ju Swjatowit. — —

A rapakow so nahna na Lubin
Kaž djasow z bjezdna walił by so ścin.
A rubachu ze złóśću do sokoła
A žónsku jako do khłódneho stoła
Pak přitłóčichu na woporny kamjeń.
Krew ze sokoła pak a z jeje ramjeń
So liješe a čeŕwjeny a čorny
Błyšć pjerjow přebarbješe sokołej
A mjehki mošk a somot płašćej jej'
A na dnje zrudne zelene te dorny.
A běły sokoł wupřestrěł bě w klinje
So mortwy wěšćeŕcy tam na Lubinje.
A sowy jako dźěći zapłaknychu
Pře zahinjenu serbskich horow pychu —

Wot Čornoboha hač do Běłoboha
Pak ržeše z błyskom hrimotanjow dróha ...

TEI Kódovani e textu

Časť textu v kódovaní TEI

```
▼<p facs="#facs_11_r_8_1">
  ▼<lg>
    ▼<l facs="#facs_11_r_8_1l1">
       <hi rend="bold:true;">W pruzy njebjeskej.</hi>
     </l>
     <l facs="#facs_11_r_8_1l2">Cyklus basni Jakuba Ćišinskeho.</l>
   </lg>
 </p>
```

Export

Časť kódovania textu, s. 11, časopis Lužica

```
<l facs="#facs_11_r_8_l1">
  <hi rend="bold:true;">W pruzy njebjeskej.</hi>
</l>
<l facs="#facs_11_r_8_l2">Cyklus basni Jakuba Ćišinskeho.</l>
</lg>
</p>
<p facs="#facs_11_r_9_1">
<lg>
<l facs="#facs_11_r_9_l1">Lubinska sonina.</l>
<l facs="#facs_11_r_9_l2">Kaž we snje Lubin stoji z wječora - </l>
<l facs="#facs_11_r_9_l3">A njebjo zamróčene woko ma –</l>
<l facs="#facs_11_r_9_l4">Ze črjódami kaž k smjerći wokoł' hory</l>
<l facs="#facs_11_r_9_l5">A z na trach krakanjom so honja wróny:</l>
<l facs="#facs_11_r_9_l6">A delka płakaju na prózdne dwory</l>
<l facs="#facs_11_r_9_l7">A noša styšć přez štomow stare króny.</l>
<l facs="#facs_11_r_9_l8">Po polach stona wokoł' hory delka,</l>
<l facs="#facs_11_r_9_l9">Hdźež lóšt bě był a było lěhwo wjelka.</l>
<l facs="#facs_11_r_9_l10">A z hłubin na wjeŕch zazdychuja styski,</l>
<l facs="#facs_11_r_9_l11">Hdźež ržałe běchu modlitwy a wyski...</l>
<l facs="#facs_11_r_9_l12">Jow serbskich ludow zabitych je row</l>
<l facs="#facs_11_r_9_l13">A žałosćacych wutrobow nět skhow.</l>
<l facs="#facs_11_r_9_l14">Přez jědle, šmrěki wichor wrjeskota,</l>
<l facs="#facs_11_r_9_l15">Kaž bychu wojownicy přez lěs hnali;</l>
<l facs="#facs_11_r_9_l16">A wjeŕški koći, k zemi praskota,</l>
<l facs="#facs_11_r_9_l17">Kaž z mječom do škitow sej bychu prali ...</l>
<l facs="#facs_11_r_9_l18">A zdobom kaž by wichor wotdychnył,</l>
<l facs="#facs_11_r_9_l19">'noz ćeńki płač by słyšeć z rowow był.</l>
<l facs="#facs_11_r_9_l20">A swětło ze skały na wjeŕchu prasny</l>
<l facs="#facs_11_r_9_l21">A wobraz w nim so zjewi čisty, jasny:</l>
<l facs="#facs_11_r_9_l22">Tam žónska stoji stawow hoberskich,</l>
<l facs="#facs_11_r_9_l23">Ta poslednja ze žonow wěšćeŕskich.</l>
<l facs="#facs_11_r_9_l24">Płasć wot ramjenjow po postawje pada</l>
<l facs="#facs_11_r_9_l25">Kaž z moška mjehki tak a běły tkany;</l>
<l facs="#facs_11_r_9_l26">A z włosow błyšć kaž z pruhow słónčnych rjany</
<l facs="#facs_11_r_9_l27">Ju z koškom spěšnym majka po khribjeće;</l>
<l facs="#facs_11_r_9_l28">A króna dejmanty do złota pleće.</l>
<l facs="#facs_11_r_9_l29">A woheń z kuzłarskeju wočow hlada,</l>
<l facs="#facs_11_r_9_l30">Kiž sylzy z njeju do rubinow twori;</l>
<l facs="#facs_11_r_9_l31">A ze rta jako płomjo ji so hori.</l>
<l facs="#facs_11_r_9_l32">Nad hłowu běły sokoł křidle šěri</l>
```

Iniciatíva kódovania textu (TEI)

- skratka TEI = *Text Encoding Initiative* (TEI)
- prekladáme ako *Iniciatíva kódovania textu*
- TEI je medzinárodný projekt, ktorý bol založený v roku 1987
- zameraný na pravidlá prípravy a výmeny elektronických textov pre oblasť vedeckého výskumu
- je určený aj pre širšie uplatnenie a usiluje sa uspokojovať aj široké potreby využívania v oblastiach, kde sa pracuje s jazykom
- (v oblasti *language industries*).

Podstata TEI

- Metóda TEI potvrdzuje tézu o tom, že informatické aplikácie by mali
- stavať na komplexnom texte a jeho štruktúre a nie len na jeho jednotlivých jazykových, napríklad morfologických alebo lexikálnych zložkách

Štruktúra textu

- Z metodologického hľadiska ide v TEI o *štrukturálny prístup* k textu
- Text sa rozkladá značkovaním na časti
- Na rozklad služia v TEI špeciálne prostriedky
- Kódovanie vychádza zo štandardu SGML
- Teoreticky ide o podobnú stratégiu štruktúrovania textu, akú reprezentujú formáty MARC
- Technika TEI je viazaná výhradne na texty v elektronickej forme
- Softvéry môžu túto techniku do určitej miery vykonávať súbežne s tvorbou textu v elektronickej forme

TEI ako metatext

- pri TEI sa vytvára metatextový (metadátový) obraz dokumentu avšak tento obraz je sémanticky a rozsahom v podstate identický (totožný) s originálnym textom
- Metatext TEI je bohatší práve len o značky dodané v procese kódovania textu podľa pravidiel TEI
- *MARC* záznam je výsledkom komprimácie, kondenzačnej deskripcie, v dôsledku ktorej je záznam o dokumenty rozsahom menší a sémanticky chudobnejší ako dokument
- Zo sémantického hľadiska sa kondenzáciou v markovskom systéme pretvára text dokumentu z úrovne *väčšej konkrétnosti* na úroveň *väčšej všeobecnosti.*

Metaúdaje/metadáta

- Z hľadiska informačnej práce je podstatné, že techniky rozličnej štrukturácie textov metadáta/metaúdaje ako
- TEI,
- HTML,
- SGML,
- MARC,
- XML,
- Rôzne hypertextové systémy a pod.
- Umožňujú konverziu, výmenu a vzájomné využívanie produktov kódovania

Značka a značkovanie

- Samotná myšlienka značkovania textu nie je nová.
- Slovo *značka, značkovanie* (markup) sa používalo na pomenovanie poznámok, symbolov, značiek, ktoré boli umiestnené priamo v texte a slúžili ako inštrukcia pre sadzača alebo tlačiara, ako má vysádzať alebo upraviť jednotlivé časti textu.
- Napríklad, ak bol text v predlohe podčiarknutý vlnovkou, bol to pre sadzača pokyn, aby túto časť vytlačil tučne (boldom).
- Na označovanie odsadenia od okraja, vynechanie riadkov, použitie zvláštneho fontu sa používali rôzne špeciálne značky.
- Príkladom takéhoto značkovania sú napríklad aj korektorské značky.
- Postupne sa formátovanie a tlač textov automatizovala a termín *značkovanie* sa postupne rozšíril a pokrýva všetky druhy špeciálnych značiek, ktoré sa vkladajú do elektronických textov a slúžia na riadenie úpravy (formátovania), tlače alebo iného spracovania.

Kódovanie = značkovanie

- Termín *kódovanie* je v danom kontexte synonymom termínu *značkovanie*

SGML ISO 8879

- Zásady a odporúčania, ktoré sú rozpracované v rámci TEI a v Pravidlách TEI sú vypracované v súlade s normou ISO 8879 (1986),
- I*SO 8879:1986 : Information processing - Text and office systems - Standard Generalized Markup Language (SGML),* [Geneva] : ISO, 1986.
- Norma definuje *Štandardný všeobecný značkovací jazyk.*
- TEI je aplikáciou normy SGML v oblasti spracovania textov, podobne ako je HTML aplikáciou SGML v prostredí WWW (domovské stránky).

SGML

- SGML je medzinárodný štandard na popis a značkovanie elektronických textov.
- Presnejšie, je to metajazyk, formálno-popisný jazyk, ktorý v danom prípade slúži ako jazyk na *značkovanie.*
- Rozumie sa ním *explicitná* (jasná, jednoznačná, zreteľná, viditeľná) interpretácia textu.
- Nejde teda o implicitnú vlastnosť textu, ale o elementy, ktoré sa do textu dostávajú *dodatočne*, *zvonku* v procese, ktorý nie je identický s procesom tvorby textu.

Bežné kódovanie

- V bežnom zmysle sú texty kódované napríklad tak, že obsahujú:

1. interpunkčné znaky (bodka, čiarka, bodkočiarka, výkričník ...),
2. veľké začiatočné písmená,
3. rozmiestnenie písmen na strane,
4. medzery medzi slovami, odsekmi atd.

Význam bežného značkovania

- Značkovanie pomáha čitateľovi určiť začiatky a konce slov, identifikovať väčšie štrukturálne celky textu, ako sú nadpisy, odstavce, vety a pod.
- Keď sa text kóduje pre počítačové spracovanie, ide v podstate o transformáciu lineárneho textu pomocou značiek
- V tejto transformácii sa do textu pridávajú explicitné značky a text sa formálne delinearizuje, rozkladá, štrukturuje.

Značkovací jazyk

- pri zápise textu sa používa *značkovací jazyk*,
- je to súbor značkovacích konvencií (pravidiel), ktoré sa spolu používajú na kódovanie textov
- značkovací jazyk (teda štandard SGML, napr. v aplikáciinTEI)musí špecifikovať,
  1. ktoré značky sa majú používať,
  2. ako sa majú používať,
  3. ako sa oddeľujú od textu a
  4. čo ktorá značka znamená

Princípy SGML

- Všeobecné zásady pre značkovací jazyk poskytuje norma SGML.
- Je založená na troch charakteristikách, ktorými sa koncepcia SGML odlišuje od iných značkovacích jazykov:

1. kladie väčší dôraz na *popisné značkovanie* ako na spracovateľské značkovanie; (značky „pridáva“ počítač)
2. pracuje s koncepciou typu dokumentu (DTD);
3. je nezávislá na systéme reprezentácie písma, ktorým je text napísaný.

SGML parsery

- softvéry, ktoré sú schopné podporovať tvorbu, hodnotenie a spracovanie dokumentov SGML
- ťažiskom týchto softvérov je *analyzátor syntaxe* (SGML parser).
- je to časť softveru, ktorá dokáže definovať typ dokumentu (DTD) a generovať z neho softvérový systém, ktorý je schopný identifikovať typ dokumentu a vyvolať procedúry pre daný typ dokumentu

Význam parserov

- existujú softvéry, ktoré sú schopné na základe syntakticej analýzy zistiť novú kánonickú formu dokumentu a formátovať dokument podľa používateľských špecifikácií.
- Takúto formu môžu použiť ďalšie časti softveru, ktoré sú viac alebo menej spojené s parserom a uskutočňovať ďalšie funkcie, ako je napríklad štruktúrované editovanie, formátovanie a manažment databázy

DTD

- *Document type definition – DTD*
- hlavným a prvým krokom textovej analýzy pomocou softvéru je určiť typ dokumentu
- Predpokladá sa, že dokument patrí k nejakému typu dokumentov.
- Typ dokumentu sa môže formálne identifikovať podľa toho, aké zložky, časti obsahuje, z čoho sa skladá.
- Čiže, analýzou štruktúry textu je možné zistiť, o aký typ dokumentu sa jedná.
- Napríklad dizertácia (diplomovka) má štandardne meno autora, názov, predhovor, abstrakt, obsah, kapitoly, ilustrácie, zoznam bibliografických odkazov a pod.

Význam parserov

- Ak ide o známy typ dokumentu, je možné použiť parser, čiže syntaktický analyzátor, ktorý dokáže zistiť, či dokument obsahuje všetky potrebné časti a či sú elementy správne usporiadané
- Dôležité je, že rôzne dokumenty toho istého typu sa dajú spracúvať rovnakým spôsobom
- Program môže byť schopný vyčleniť poznatky, ktoré sú ukryté v dokumente (knowledge encapsulated in the document structure information) a pomáhať tak používateľovi ako inteligentný pomocník

Štruktúra SGML

- Štruktúru SGML tvorí jednotný konzistentný mechanizmus na značkovanie alebo identifikáciu textových štrukturálnych jednotiek, ktoré sú definované v SGML.
- V struktúre sa tiež definuje, ako kombinovať štrukturálne prvky, ktoré sa vyskytujú v texte.

Štruktúra SGML

- Štruktúru SGML tvoria:

1. elementy

2. Atribúty draft | revised | published)

3. konektory

4. entity

Elementy

- **elementy - textové jednotky ako štrukturálne zložky;**

napr.: <anthology>, <poem><title>The SICK ROSE</title>

Atribúty

draft | revised | published)

Konektory

- **konektory - značky na spájanie viacerých komponentov, ako napríklad čiarka, zvislá čiara, ampersand; napr.:**

(TITLE?, STANZA+), <!ELEMENT (line | line1 | line2) O O (#PCDATA) >

Entity

- **entity - označená časť dokumentu, ktorá predstavuje zámer štruktúrovania; je to určitý reťazec znakov alebo textový celok; napr.:**

<!ENTITY tei "Text Encoding Initiative">

Príklad

Báseň The SICK ROSE od Williama Blakea z antológie Songs of innocence and experience (1794).

```
<anthology>
    <poem><title>The SICK ROSE</title>
      <stanza>
        <line>O Rose thou art sick.</line>
        <line>The invisible worm,</line>
        <line>That flies in the night</line>
        <line>In the howling storm:</line>
      </stanza>
      <stanza>
        <line>Has found out thy bed</line>
        <line>Of crimson joy:</line>
        <line>And his dark secret love</line>
        <line>Does thy life destroy.</line>
      </stanza>
    </poem>
        <!-- more poems go here  -->
</anthology>
```

Báseň

Je to jednoduchý model štruktúry textu SGML/TEI

Identifikovaný je DTD - typ dokumentu (anthology), čo znamená, že v dokumente sú napr zbierka - básne.

<anthology> začiatok označenia typu dokumentu

</anthology> koniec označenia typu dokumentu

<poem> začiatok básne (poémy)

</poem> koniec básne (poémy)

Potom môže nasledovať ďalšia báseň (poem)…

- Každá báseň v antológii je jedným elementom báseň (poem), v ktorom je

názov (title),

<title> začiatok titulu </title> je koniec titulu

strofa (stanza),

<stanza> začiatok strofy </stanza> je koniec strofy

riadok (line), predstavujúci verš.

<line> začiatok riadku/verša</line> je koniec riadku verša