

Iniciatíva kódovania textu (TEI)

Prof. PhDr. Dušan Katuščák, PhD.
Slezská univerzita, FPF-ÚBK Opava

Server Overview Layout 2

Logout dusankatuscak@gmail.com

Document... Find

Document Mana User Manager

Versions Jobs

Recent document User activity

Collections: INT_West slavonic Col-ID

Documents Model Data

1-21 / 21

Title

- 4... Copy of HTR Validation Set 'S.
- 4... Copy of HTR Train Set 'Sriptc
- 4... TRAINING_VALIDATION_SET_S
- 4... Copy of HTR Train Set 'FRAKT.
- 2... Slovenske narodnje novini 18
- 1... Šlabikář 1872
- 5... Ewanjelický Šlabikář wypraco.
- 0... TRAINING_VALIDATION_SET_I
- 0... TRAINING_VALIDATION_SET_I
- 5... **Luzica_PDF**
- 4... Ioann. Amos Comenii Orbis Pi
- 5... PALUGYAY
- 5... JITRENKA.JPG
- 5... Cirkew Ewanjelicko -Lutheráns
- 9... TRAINING_VALIDATION_SET_F
- 9... Jánošík
- 9... Noviny OBZOR_1866.pdf
- 3... Opavský Besedník (2)
- 4... TRAINING_VALIDATION_SET_...
- 3... Moravské noviny

100 Filter



- 1-1 Wukhadža srjedź kóždoho
- 1-2 měsaca w Budyšinje a hodži
- 1-3 so skazać pak na póstach,
- 1-4 pak pola administratora a
- 1-5 w expediciji „Serb. Now.“
- 2-1 Dleći lětnje w Němskej:

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:

Filename pattern

- pageNr + filename
- filename (warning: filenames must be unique for document)
- docId + pageNr + pageId

Pattern:

Placeholder: \${docId}, \${filename}, \${pageId}, \${pageNr}, \${filekey}

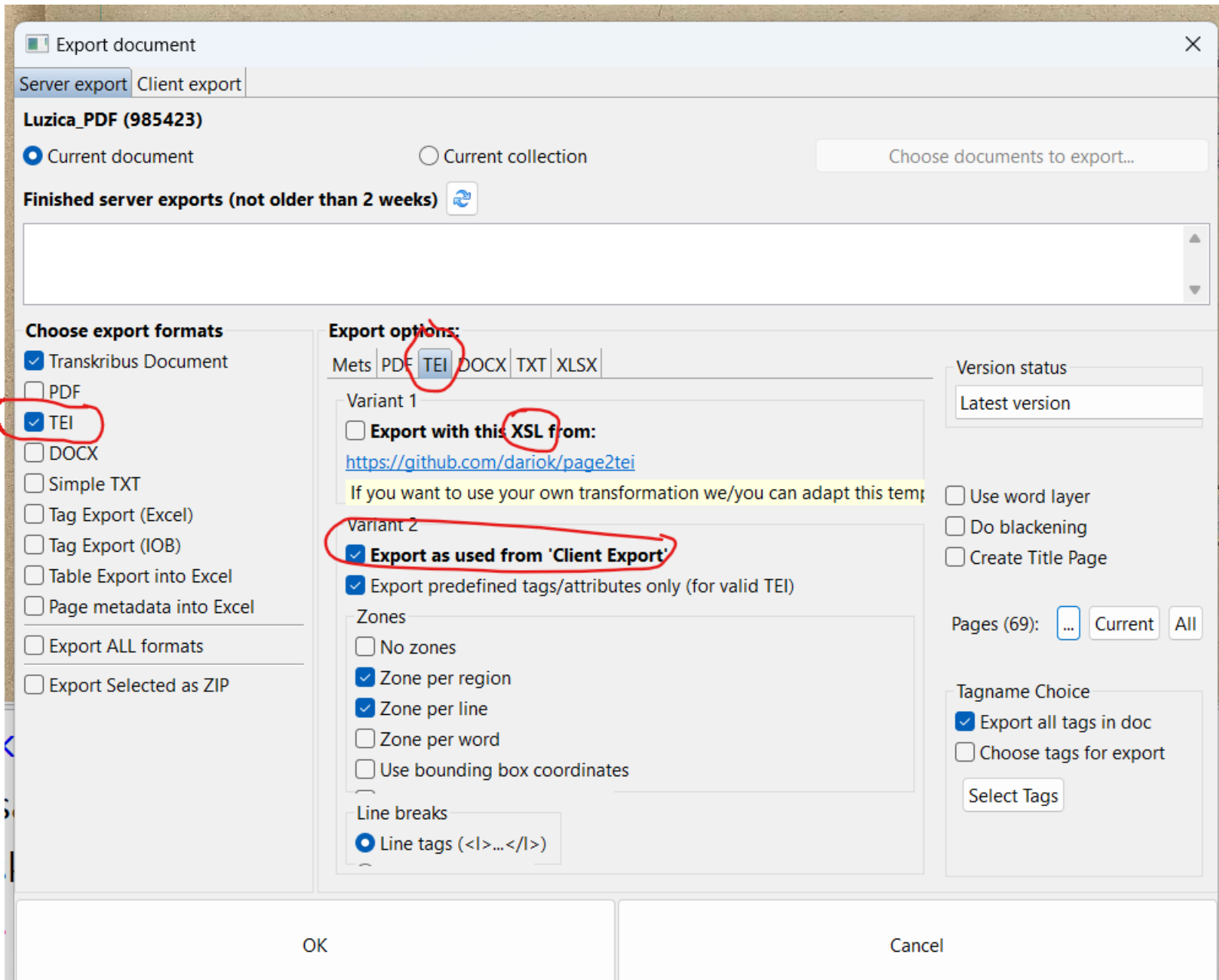
Version status

- Use word layer
- Do blackening
- Create Title Page

Pages (69):

OK

Cancel



Export document complete Doručené x



transkribus@uibk.ac.at
komu: mne ▾

22:08 (pred 0 minútami) ☆ ↶ ⋮

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 srjeđó kóždého
mésaca w Budyšinje a hodći
so skazać pak na póstaeh,
pak pola administratora
w expediciji „Serb. Now.“

ŁUŽICA.

Plaći lótnje w Němskej:
3 hriwny, w Rakuko-Wuhé-
skej: 4 króny, we Wukhodźe:
2 rublje, w Zapadźe: 4 franki.

Měsačnik za zabawu a powučenje. Zhromadny časopis hornjo- a delnołužiskich Serbow.

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

Číslo 1.

Januar 1909.

Lětnik 28.

W pruzy njebjeskej.

Cyklus basni Jakuba Čišinskeho.

Lubinska sonina.

Kaž we snje Lubin stoji z wječora —
A njebjo zamróčene woko ma —
Ze črjódami kaž k smjeróci wokol' hory
A z na trach krakanjom so honja wróny:
A delka plakaju na prózdne dwory
A noša styšć přez štomow stare króny.

Po polach stona wokol' hory delka,
Hdžež lóšt bě byt a było lěhwo wjelka.
A z hlubin na wječeh zazdychuja styski,
Hdžež ržale běchu modlitwy a wyski ...

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

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

A zdobom kaž by wichor wotdychnył,
'nož čěňki plač by słyšeć z rowow byl. —

A swěto ze skały na wječehu prasny
A wobraz w nim so zjewi čisty, jasny:
Tam žónska stoji stawow hoberaskich,
Ta poslednja ze žonow wěšćešskich.

Plašć wot ramjenjow po postawje pada
Kaž z moška mjehki tak a běly tkany;
A z włosow blyšć kaž z pruhow slóněnych rjany
Ju z koškom spěšnym majka po khibjeće;
A króna dejmanty do złota pleće.
A woheń z kuzlařskej wočow hlada,
Klěž sylzy z njeju do rubinow twori;
A ze rta jako plomjo ji so hori,
Nad hłowu běly sokol křidle šěri

A z wótrym wokom do dala so měri;
A z hlówy stawa jemu pjerjow blyšć,
Zo dótknyć njemohl so ni strach ni styšć.

A žónska měša w kótle džiwny war,
Pak khwalba ze rta šepce ji pak swar.
A z hlósom k njebjesam so woła wona
A z hórkim plačom čiši styskne hrona.
A wokol' kóta z džiwnej nohu khodźi,
Prut złoty do kóta, pak z kóta wodźi.
A jako sokol woko wupina
Do štyrjoch stronow dele z Lubina ...

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

A rapakow so nahna na Lubin
Kaž djasow z bjezdna walil by so sčín.
A rubachu ze zlósc do sokola
A žónsku jako do khlódného stola
Pak přitlóšiču na woporny kamjeń.
Krew ze sokola pak a z jeje ramjeń
So liješe a čerwjeny a čorny
Blyšć pjerjow přebarbješe sokolej
A mjehki mošk a somot plašćej jej!
A na dnje zrudne zelene te dorny.
A běly sokol wupřestrěl bě w klinje
So mortwy wěšćečy tam na Lubinje.
A sowy jako džěci zaplaknychu
Pře zahinjenu serbskich horow pychu —
Wot Čornoboha hač do Běłoboha
Pak ržeše z blyskom hrimotanjow dróha ...

TEI

Kódování e textu

Část textu v
kódování TEI

```
▼<p facs="#facs_11_r_8_1">  
  ▼<lg>  
    ▼<l facs="#facs_11_r_8_111">  
      <hi rend="bold:true;">W pruzy njobjeskej.</hi>  
    </l>  
    <l facs="#facs_11_r_8_112">Cyklus basni Jakuba Čišinskeho.</l>  
  </lg>  
</p>
```

Export

Časť kódovania textu, s. 11, časopis Lužica

```
...o.
▼<l facs="#facs_11_r_8_111">
  <hi rend="bold:true;">W pruzy njebjeskej.</hi>
</l>
  <l facs="#facs_11_r_8_112">Cyklus basni Jakuba Čišinskeho.</l>
</lg>
</p>
▼<p facs="#facs_11_r_9_1">
  ▼<lg>
    <l facs="#facs_11_r_9_111">Lubinska sonina.</l>
    <l facs="#facs_11_r_9_112">Kaž we snje Lubin stoji z wječora - </l>
    <l facs="#facs_11_r_9_113">A njebjo zamróčene woko ma -</l>
    <l facs="#facs_11_r_9_114">Ze črjódami kaž k smjerći wokoł' hory</l>
    <l facs="#facs_11_r_9_115">A z na trach krakanjom so honja wróny:</l>
    <l facs="#facs_11_r_9_116">A delka pĹakaju na prózdne dwory</l>
    <l facs="#facs_11_r_9_117">A noša styšč přez štomow stare króny.</l>
    <l facs="#facs_11_r_9_118">Po polach stona wokoł' hory delka,</l>
    <l facs="#facs_11_r_9_119">Hdžež lóšt bě był a było lěhwo wjelka.</l>
    <l facs="#facs_11_r_9_1110">A z hłubin na wjerch zazdychuja styski,</l>
    <l facs="#facs_11_r_9_1111">Hdžež ržałe běchu modlitwy a wyski...</l>
    <l facs="#facs_11_r_9_1112">Jow serbskich ludow zabitych je row</l>
    <l facs="#facs_11_r_9_1113">A žalosćacych wutrobow nět skhow.</l>
    <l facs="#facs_11_r_9_1114">Přez jědle, šmrěki wichor wrjeskota,</l>
    <l facs="#facs_11_r_9_1115">Kaž bychu wojownicy přez lěs hnali;</l>
    <l facs="#facs_11_r_9_1116">A wjerški koći, k zemi praskota,</l>
    <l facs="#facs_11_r_9_1117">Kaž z mječom do škitow sej bychu prali ...</l>
    <l facs="#facs_11_r_9_1118">A zdobom kaž by wichor wotdychnył,</l>
    <l facs="#facs_11_r_9_1119">'noz ćeńki pĹač by slyšeć z rowow był.</l>
    <l facs="#facs_11_r_9_1120">A swětło ze skały na wjerchu prasny</l>
    <l facs="#facs_11_r_9_1121">A wobraz w nim so zjewi čisty, jasny:</l>
    <l facs="#facs_11_r_9_1122">Tam žónska stoji stawow hoberskich,</l>
    <l facs="#facs_11_r_9_1123">Ta poslednja ze žonow wěšćeńskich.</l>
    <l facs="#facs_11_r_9_1124">PĹasć wot ramjenjow po postawje pada</l>
    <l facs="#facs_11_r_9_1125">Kaž z moška mjehki tak a běly tkany;</l>
    <l facs="#facs_11_r_9_1126">A z włosow blyšč kaž z pruhow slóńčnych rjany</l>
    <l facs="#facs_11_r_9_1127">Ju z koškom spěšnym majka po khríbjeće;</l>
    <l facs="#facs_11_r_9_1128">A króna dejmanty do złota pleće.</l>
    <l facs="#facs_11_r_9_1129">A woheń z kuzłarskeju wočow hlada,</l>
    <l facs="#facs_11_r_9_1130">Kiž sylzy z njeju do rubinow tworí;</l>
    <l facs="#facs_11_r_9_1131">A ze rta jako pĹomjo ji so horí.</l>
    <l facs="#facs_11_r_9_1132">Nad hłowu běly sokoł křidle šeri</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 morfológických alebo lexikálnych zložkách

Štruktúra textu

- Z metodologického hľadiska ide v TEI o *štrukturálny prístup* k textu
- Text sa rozkladá značkováním na časti
- Na rozklad slúžia 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 štruktúrácie 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),
- *ISO 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 začiatkové 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 štruktúrne 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 aplikácii TEI) 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 parseery

- 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 syntaktickej 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 štruktúre 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 básne (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