Iniciatíva kódovania textu (TEI)

Dušan Katuščák

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
- vychádzajú 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 metatextvový 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
- *markovský* 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

- 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

- 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.*
- Využíva tiež ISO 646, ktorá definuje sedembitovú znakovú sadu
- 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:
- interpunkčná znaky (bodka, čiarka, bodkočiarka, výkričník ...),
- veľké začiatočné písmená,
- rozmiestnenie písmen na strane,
- 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 transkribciu lineárneho textu rukopisu, čiže o transformáciu.
- 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 musí špecifikovať, ktoré značky sa majú používať, ako sa majú používať, ako sa oddeľujú od textu a 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:
  - kladie väčší dôraz na *popisné značkovanie* ako na spracovateľské značkovanie;
  - pracuje s koncepciou typu dokumentu (DTD);
  - je nezávislá na systéme reprezentácie písma, ktorým je text napísaný.

SGML parsery

- softvery, 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
- Pkaždý redpokladá 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

- Štruktúrovaný editor je druhom inteligentného textového procesora
- Dokáže použiť informáciu extrahovanú zo spracovaného DTD a ukázať používateľovi informáciu o tom, ktoré prvky sú potrebné v jednotlivých fázach tvorby dokumentu. Dokáže tiež mimoriadne zjednodušiť prípravu dokumentu napríklad tak, že automaticky vkladá do textu tagy.

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:
- **elementy**
- **atribúty** (draft | revised | published)
- **konektory**
- **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.
- Identifikovaný je typ dokumentu (anthology), čo znamená, že v dokumente sú básne.
- Každá báseň je jedným elementom, v ktorom je názov (title), strofa (stanza), riadok (line), predstavujúci verš.