Masarykova univerzita, Fakulta informatiky

Digitální knihovny PV070

Text Encoding Initiative – TEI

http://www.tei-c.org/index.xml

Antonín Polčák, 4. ročník, 4. 12. 2013

Co je to TEI

TEI je neziskové společenství, které vytváří a udržuje standard pro reprezentaci textů v digitální podobě. Je složené z akademických institucí, výzkumných projektů a vědců z celého světa. Hlavním posláním je vytvoření sady směrnic či pokynů, které specifikují metody kódování pro strojově čitelné texty, a to především v lingvistice, humanitních a společenských vědách. Kromě směrnic TEI poskytuje i řadu vzdělávacích workshopů pro výuku TEI a rovněž software, který je pro TEI vyvinut nebo přizpůsoben.

Sdružení TEI sestává z technické rady, podpůrné TEI komunity a představenstva. Technická rada se stará o organizaci a technický rozvoj TEI směrnic. Svolává tedy pracovní skupiny pro řešení konkrétních projektů. Představenstvo poskytuje strategický směr a dohled na rozpočet. Dále má na starost marketing, koordinaci shromažďování veřejných prostředků a členství. Úspěch TEI je odvislý od aktivní účasti členů komunity a uživatelů.

Historie

TEI byla založena v roce 1987, a to k rozvoji, údržbě a propagaci hardwarově a softwarově nezávislých metod kódování údajů v elektronické podobě.

Počátky

V době, kdy byla TEI založena, se vědecké projekty a knihovny snažily využívat digitální technologie. Potýkaly se však s velkou překážkou – vytvoření udržitelných a sdílených archivů a nástrojů, tedy systémy, jež by zastupovaly textové materiály. Tyto systémy byly navzájem často nekompatibilní, obvykle byly špatně navržené a počet těchto systémů se rychle zvyšoval. Tato situace ovšem zpomalovala růst využití plného potenciálu počítačů, a to kvůli překážkám v přístupu, vznikem nových problémů s uchováním dat (díky čemuž bylo jejich sdílení velmi obtížné) a také kvůli nepraktickým vývojem společných nástrojů. Za účelem řešení těchto problémů se systémy, byla v listopadu 1987 na Vassar College svolána schůzka. Sešla se různorodá skupina vědců z mnoha různých oborů. Byla zde zastoupena odborná veřejnost, knihovny, archivy a projekty z řady evropských států, Severní Ameriky a Asie. Organizace prací na vývoji směrnic TEI byla prováděna třemi sponzorskými organizacemi: The Association for Computers in the Humanities, the Asocitation for Literary and Linguistic Computing a the Association for Computational Linguistics. Řídící výbor byl tvořen zástupci těchto asociací. Pracovat začali dva vybraní editoři, ale do konce roku 1989 bylo do prací přímo zapojeno více než 50 vědců a podpora rychle rostla. Počáteční fáze vyústila ve vydání první verze směrnic, známé jako P1, v červnu 1990. Druhá fáze byla zveřejněna v průběhu let 1990-1993. Přibylo přitom 15 pracovních skupin, byly provedeny revize a různá rozšíření. V květnu 1994 byla vydána první oficiální verze směrnic – P3. Zatímco se na projektu stále usilovně pracovalo a poměrně rychle se měnily cíle i návrhy, byly P3 směrnice přijaty řadou projektů. Postupně byly prováděny různé vzdělávací workshopy a semináře pro podporu dalšího rozvoje. Základní skupina vědců se postupně rozšířila až na 200.

TEI konsorcium

V lednu roku 1999 předložily univerzity ve Virginii a norském Bergenu návrh výkonnému výboru TEI pro vytvoření mezinárodní organizace – TEI konsorcia, které by zachovávalo rozvoj a podporu TEI. Návrh byl přijat a krátce nato byly přidány další dvě instituce s dlouholetými vazbami na TEI – Brown University a Oxford University. V roce 2000 byla dohoda o vytvoření konsorcia podepsána. Po vzniku TEI konsorcia bylo prioritou vydání XML verze TEI směrnic, které by uživatelům umožňovaly pracovat

s nově vznikajícími nástroji XML. V červnu 2002 tak vznikla nová verze – P4. Byla to však pouze XML verze P3 s menšími opravami chyb. Hned se tak začalo pracovat na nové verzi, která by obsahovala důkladnější opravy chyb, zahrnujícím navíc připomínky veřejnosti a vývoj v klíčových oblastech, jako je grafika, kódování znaků, jazyk směrnic apod. P5 verze směrnic byla vydána v listopadu 2007.

Dopad projektu TEI byl obrovský. TEI je dnes mezinárodně uznávaným nástrojem pro dlouhodobé uchovávání elektronických dat, a jako prostředek podporující efektivní využívání těchto údajů. TEI výrazně napomohla předložení našeho kulturního dědictví současnému, internetem propojenému světu a bude tak plně k dispozici studentům, vědcům a široké veřejnosti.

Směrnice TEI

Směrnice definují a popisují značkovací jazyk, který je určen pro reprezentaci strukturálních, interpretačních a pojmových vlastností textů. Směrnice jsou vyjádřeny jako XML schéma, jež lze přizpůsobit či rozšířit. Dále obsahují také podrobnou dokumentaci. Pokyny TEI jsou přitom publikovány pod open-source licencí, což znamená, že jejich kód je veřejně přístupný.

Aktuální verze

Nejnovější verze TEI směrnic je znám pod označením **P5**. Vydána byla v listopadu 2007 a v šestiměsíčním cyklu je pravidelně aktualizována.

TEI směrnice jsou v současné době přeloženy z angličtiny do několika světových jazyků. Momentálně jsou přístupné v čínštině, francouzštině, němčině, italštině, japonštině, korejštině a španělštině.

TEI dále provozuje internetové stránky SourceForge, kde je řízen vývoj a distribuce aktuálních verzí směrnic. Je možno zde stáhnout jejich zdrojové soubory a související materiály. Zároveň jsou v archivu TEI přístupny i starší verze a příslušné materiály těchto verzí.

Uživatelům je na oficiálních stránkách TEI přístupná také webová aplikace Roma. Ta umožňuje vygenerovat schémata a dokumentaci, které jsou kompatibilní se směrnicemi P5. Dokumentaci je možno tvořit i v jiných světových jazycích. Roma je nástupce webového nástroje TEI Pizza Chef, který vytvářel DTD na míru, jež byly kompatibilní s P4.

TEI dále udržuje knihovnu XSL stylů, které umí konvertovat TEI XML soubory do HTML, LaTeXu nebo XSL:FO dokumentů.

Dokumentace směrnic

Pro názornost uvedu krátký výtažek z dokumentace ke směrnicím P5.

Takto může vypadat elektronicky zpracovaný text, bez zpracování pomocí TEI:

Review

Die Leiden des jungen Werther[1] is an *exceptionally* good example of a book full of *Weltschmerz*.

Obrázek 1 Ukázka nezpracovaného elektronického textu

- problémem zde mohou být mimo jiné čísla stránek, uvozovky a apostrofy, dělení slov, konce řádků atd.

A takto vypadá příklad po použití kódování podle směrnic TEI P5:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Review: an electronic transcription</title>
         </titleStmt>
         <publicationStmt>
            <p>Published as an example for the Introduction module of TBE.
            </p>
         </publicationStmt>
         <sourceDesc>
            <p>No source: born digital.</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <head>Review</head>
         <p>
            <title>Die Leiden des jungen Werther</title>
            <note place="foot">by <name>Goethe</name></note>
         is an
            <emph>exceptionally</emph>
         good example of a book full of <term>Weltschmerz</term>.</p>
      </body>
   </text>
</TEI>
```

Obrázek 2 Ukázka zpracovaného elektronického textu

- odstavce a kapitoly jsou označeny, apostrofy jsou od uvozovek odlišeny, stránky jsou označeny značkami <pb /> apod.

Struktura TEI textu

Všechny texty, které strukturně vyhovují TEI, obsahují:

- *TEI hlavičku* – značeno prvkem <teiHeader>. Poskytuje podobné informace, jako titulní stránka tištěného dokumentu. Může mít až čtyři části (popis textu, popis způsobu kódování apod.)
- *Správný přepis textu* – ohraničen elementem <text>. Tento blok může být buď jednotný (jeden díl) nebo složený (více dílů). V obou případech může obsahovat přední část <front> (záhlaví, název stránky, předmluvy atd.) a zadní část <back> (dodatky apod.). Mezi nimi je pak hlavní obsah textu <body>.

Tyto dva prvky (hlavička a přepis textu) jsou formovány do jednoho <TEI> elementu, který musí být deklarován v rámci TEI namespace1.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
<!-- [ TEI Header information ]  -->
 </teiHeader>
 <text>
  <front>
<!-- [ front matter ... ] -->
  </front>
  <body>
<!-- [ body of text ... ]  -->
  </body>
  <back>
<!--  [ back matter ...  ] -->
  </back>
 </text>
</TEI>
```

Obrázek 3 Ukázka jednotného bloku textu

V případě složeného bloku by bylo bloků <text> více a byly by všechny dohromady obklopeny elementem <group>.

Tělo textu může být děleno do kapitol, sekcí, podsekcí atd. Každý odstavec je označován elementem <p>. Element obsahuje členění přední či zadní částí a těla. Může být dále specifikován pomocí různých atributů. Např. xml:id se používá jako unikátní identifikátor dané sekce.

Tímto způsobem se postupuje hlouběji a jsou specifikovány mnohé elementy k rozlišení různých částí textu.

Dále je také možno využívat ukazatelů a odkazů na jiné místo v dokumentu nebo v jiném dokumentu. Datum a čas, čísla, seznamy, citace apod. mají také své speciální elementy, aby mohli být správně kódovány.

Elektronický text je tak ve výsledku doplněn bohatým množstvím elementů, které mu dodají mnohé možnosti, které jsou vhodné např. pro vyhledávání konkrétních parametrů textu.

Porovnání metadatových schémat MODS, MARC, Dublin Core a TEI

Závěrem bych chtěl porovnat základní parametry několika schémat s TEI.

MODS (Metadata Object Description Schema) je kompromisem mezi jednoduchým formátem Dublin Core a specializovaným MARC. Při vývoji byl kladen důraz na snížení ztrátovosti při převádění z a do jiných formátů.

MARC (Machine-Readable Cataloging) je rodina formátů a standardů určených pro ukládání a přenos bibliografických informací. MARC se postupně rozšířil do mnoha knihoven po celém světě. Základ přitom zůstal stejný, ale vznikly různé národní mutace (USMARC, UKMARC atd.). Mezi sebou však nebyly dostatečně kompatibilní a tak byl vytvořen mezinárodní formát MARC 21, který se inspiroval americkou a kanadskou formou a ve většině zemí, včetně České republiky nahradil stávající formáty.

Dublin Core je obecnější sada metadat. Je to spíše standard, který definuje pouze značky a jejich vyplnění. Konkrétní zápis pak závisí na uživateli.

TEI je mezinárodní projekt, jenž je zaměřen na vývoj standardizovaných DTD vhodných pro uchovávání a výměnu tištěných textových dokumentů v digitální podobě. Přispěla k vytvoření specializovaných DTD pro jednotlivá literární odvětví.

Struktury

MODS nezahrnuje sice kompletní set prvků MARC, ale jeho obsahem jsou i nové prvky. Přibyly i další atributy a zlepšila se převoditelnost dat. MODS sestává ze souboru bibliografických prvků, použít jej lze však také v jiných institucích. MODS je založeno na jazyce XML. Poslední verze obsahuje 20 základní prvků, které se dále dělí na podprvky. K jednotlivým prvkům i podprvkům je možno přiřadit atributy. Prvky i podprvky lze opakovat (musí však dodržovat přesné pořadí). Atributy takovéto pořadí nemají, ale nedají se opakovat.

MARC jako základ zápisu používá tři prvky:

- Strukturu záznamu – tříciferný kód, definuje význam záznamu
- Označení obsahu – třímístné číselné kódy charakterizující údajové prvky a podporující manipulaci s údaji a interpretaci obsahu
- Obsah záznamu – obsah jednotlivých polí definují různé standardy, jedná se o jednoduchý textový řetězec

Dublin Core – má dvě úrovně: základní a vyšší. Základní obsahuje 15 nepovinných prvků, které nemají pevné pořadí a lze je opakovat. Vyšší úroveň obsahuje navíc tři prvky a umožňuje využívat kvalifikátory, které upřesňují sémantiku prvku či hodnoty.

TEI definuje velké množství SGML značek, které jsou sdružovány do DTD fragmentů, což jsou sady značek. Každá takováto sada obsahuje definice jednotlivých elementů, které spolu souvisí v oblasti daného použití. Sady jsou ve výsledku obsaženy v jednom či více systémových souborech. Výsledné DTD pak na tento soubor či soubory. Toto výsledné DTD je přitom tvořeno výběrem vhodné kombinace DTD fragmentů vymezujících požadované elementy, které následně vystupují jako jeden celek.

TEI se skládá ze tří základních typů značkovacích sad:

- Centrální – obecné elementy hlavního DTD
- Základní – elementy předurčující příslušnost ke konkrétnímu typu TEI dokumentů
- Pomocné – speciální značky používané k zvláštním účelům

Vlastní zhodnocení projektu

Projekt TEI byl vytvořen za účelem lepšího zpracování elektronicky uchovávaných textů. Důvod je tedy zřejmý. TEI od začátku svého působení provedla obrovský kus práce. To dokazuje i její mezinárodní rozšíření. Popis dokumentu je díky TEI snadnější. Sady značek je totiž možno uživatelsky přizpůsobit. Vzhledem k součinnosti s XML, je navíc i kompatibilní s ostatními formáty, což je žádoucí. Pokud bude TEI pokračovat stejným tempem jako doposud, brzy se jistě dočkáme české verze TEI směrnic.

Zdroje:

[1] IDE, Nancy a Jean VÉRONIS. Text encoding initiative: background and context. Dordrecht: Kluwer Academic Publishers, c1995, 242 s. ISBN 0792337042.

[2] Metadatová schémata - srovnání. In: LÁNOVÁ, Jana. *Wikipedia: the free encyclopedia* [online]. 2012. Dostupné z: http://wiki.knihovna.cz/index.php/Metadatov%C3%A1_sch%C3%A9mata_-_srovn%C3%A1n%C3%AD

[3] TEI. *Text Encoding Initiative* [online]. Dostupné z: http://www.tei-c.org/index.xml

Dublin Core metadata


```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dcq="http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm">
<rdf:Description rdf:about="TEI - Text Encoding Initiative">
<dc:Title>TEI - Text Encoding Initiative</dc:Title>
<dc:Creator>Antonín Polčák</dc:Creator>
<dc:Description>
<rdf:Description>
<dcq:DescriptionType>abstract</dcq:DescriptionType>
<rdf:value>Neziskový projekt pro tvorbu a údržbu standardu pro reprezentaci textů v digitální
podobě</rdf:value>
</rdf:Description>
</dc:Description>
<dc:Date>
<rdf:Description>
<dcq:DateType>created</dcq:DateType>
<rdf:value>4.12.2013</rdf:value>
</rdf:Description>
</dc:Date>
<dc:Type>
<rdf:Description>
<dcq:TypeSheme>DCMIType</dcq:TypeSheme>
<rdf:value>Text</rdf:value>
</rdf:Description>
</dc:Type>
<dc:Format>
<rdf:Description>
<dcq:FormatSheme>IMT</dcq:FormatSheme>
<rdf:value>application/pdf</rdf:value>
</rdf:Description>
</dc:Format>
<dc:Format>
<rdf:Description>
<dcq:FormatType>medium</dcq:FormatType>
<rdf:value>computerFile</rdf:value>
</rdf:Description>
</dc:Format>
<dc:Language>
<rdf:Description>
<dcq:LanguageScheme>RFC3066</dcq:LanguageScheme>
<rdf:value>cze</rdf:value>
</rdf:Description>
</dc:Language>
</rdf:Description>
</rdf:RDF>
```