Zpravodaj Československého sdružení uživatelů TeXu

Marcel Svitalský

Projekt Dublin Core metadata interface

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 19 (2009), No. 1-2, 102–106

Persistent URL: http://dml.cz/dmlcz/150039

Projekt Dublin Core metadata interface

Marcel Svitalský

Abstrakt

Článek přináší překlad webové stránky popisující podrobný návrh jednoho z projektů pro GSoC 2009 pod poradenstvím TUGu. Cílem projektu měla být implementace Dublin Core Abstract Model v TEXu.

Klíčová slova: Google, Google Summer of Code 2009, TUG, TEX Users Group, metadata, Dublin Core Abstract Model, RDF, XMP, XML.

Tento dokument je návrhem projektu pro předpokládanou účast TUGu v Google Summer of Code 2009 [8]. Poradci projektu by byli Peter Flynn (University of College, Irsko) a Matthew Leingang (Courant Institute of Mathematical Sciences, New York). Jestliže vás zaujme, prosím kontaktujte mailing list TUGu o GSoC [9].

Dublin Core Metadata Initiative [2] je otevřená organizace zapojená do vývoje standardů pro interoperabilní online metadata podporujících široké rozpětí účelů a aplikačních modelů. Vyvinula abstraktní framework pro metadata [3] a několik strojově čitelných reprezentací [6] metadatových výrazů, mimo jiné v Resource Description Frameworku [13] (RDF). RDF samo má pak několik XML reprezentací [14]. Dublin Core obsahuje množinu elementů k popisu zdrojů (kupř. autor, datum vytvoření, formát), avšak jeho klíčovou vlastností je modularita: aplikace si mohou navrhnout své vlastní elementy a slovníky a přímo je použít. Zásadní také je, aby co možná nejvíce z nich, tj. jmen elementů a hodnot ve slovnících, bylo pojmenováno s užitím URI, aby se tak předešlo nejednoznačnostem.

Jedním z významných uživatelů RDF metadat je Adobe, tvůrce formátu PDF. Jejich eXtensible Metadata Platform [7] (XMP) umožňuje tvůrcům PDF dokumentů vkládat do PDF libovolná metadata. Ta jsou viditelná jak aplikacím Adobe, tak i rostoucímu množství dalších vyhledávacích a archivačních nástrojů, včetně Spotlight z Mac OS X. XMP je implementováno v XML reprezentaci RDF.

Zároveň různé nástroje pro shromažďování online bibliografických dat – jako je Zotero [20] – dávají uživatelům možnost podchytit metadata ve formátech Z39.88 [12] nebo RDF, jež jsou součástí webových stránek (nebo PDF dokumentů), přímo do jejich osobní databáze, z níž pak mohou být uložena do BibTEXu a jiných formátů odkazů pro účely okamžitého užití. Takto může PDF dokument obsahovat nejen svá vlastní metadata, nýbrž také metadata všech vnějších zdrojů, jež cituje, kupř. Seznamu literatury či klikacích odkazů.

V současnosti je již možné – například s pomocí balíku hyperxmp [16] Scotta Pakina – aby autor LATEXového dokumentu dobře obeznámený s RDF napsal

a do PDF vyprodukovaného pdfLaTeXem vložil vlastní XMP pakety. Ovšem množství i kvalita metadat vzrůstá se snadností práce s nimi, a tento způsob není právě z nejsnazších. Srovnejte si to s jednoduchými, k zapamatování i užití velmi snadnými LaTeXovými příkazy, jakými jsou `\title`, `\author`, `\date` apod. Mimo to existuje množství metadat, jako jsou struktura dokumentu či odkazy na další zdroje, jež by bylo lze vyzískat automaticky.

Klíčovými výstupy z tohoto projektu by měly být:

1. implementace Dublin Core Abstract Model v TeXu;
2. metody exportu metadat z abstraktního modelu do vnějších souborů v různých formátech, především RDF+XML, možná ale také DC-TEXT [5] a N3 [11];
3. pro pdfLaTeX automatizované vkládání XMP paketů do výsledného PDF, při defaultním minimu XMP vyjádření Z39.88 OpenURL COinS [12] polí, jak pro vlastní medatada dokumentu, tak pro všechny citované odkazy a externí hyperlinky;
4. uživatelsky přívětivé rozhraní pro vytváření metadatových výrazů;
5. chybějí-li autorské deklarace v pdfLaTeXovém dokumentu, měla by být vložena všechna metadata, jež lze detekovat automaticky;
6. metody pro autory balíků, umožňující deklarovat nové množiny metadatových elementů a slovníků, aby autoři mohli zapisovat metadata příslušných oblastí jejich zájmu. Osobně uvažuji o Learning Object Metadata [10], avšak mapování LOM do Dublin Core je problematické [17].

První ukázka použití

V LaTeXu píšící autorka chce vytvořit dokument o matematických výpočtech. Vloží tedy příkazy pro metadata do preambule dokumentu, asi takto:

```
\title{Vytváření grafů funkcí}
\author{Paní Srozuměná}
\subject{Počet}
\subject[LCSH]{Matematika -- Počet -- Diferenciální}
```

Prvý příkaz `\subject` bude čitelný pro více agentů (například pro lidi), bude však méně strukturovaný. Druhý vyhovuje Library of Congress Subject Headings, takže čtečka metadat disponující jejich slovníkem bude moci tento zdroj náležitě indexovat.

V průběhu zpracování dokumentu LaTeXem vznikne jeden nebo více RDF souborů, jež lze následně publikovat na webu. XML reprezentace výše uvedených výrazů by mohla vypadat takto:

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dcterms="http://purl.org/dc/terms/"
     xmlns:dcam="http://purl.org/dc/dcam/">
   <rdf:Description rdf:about="http://uri/for/document">
     <dcterms:creator>Paní Srozuměná</dcterms:creator>
     <dcterms:title xml:lang="cs">
       Vytváření grafů funkcí
     </dcterms:title>
     <dcterms:subject xml:lang="cs">Počet</dcterms:subject>
     <dcterms:subject>
       <dcam:memberOf rdf:resource="http://loc.gov/LCSH" />
       <rdf:value>Matematika -- Počet -- Diferenciální</rdf:value>
     </dcterms:subject>
   </rdf:Description>
</rdf:RDF>
```

Bude-li užit pdfLaTeX, vytvoří se a do PDF budou vloženy odpovídající XMP pakety. XMP paket [19] vypadá asi nějak takto:

```
<?xpacket begin="&#xFEFF;" id="W5M0MpCehiHzreSzNTczkc9d" ?>
   <x:xmpmeta xmlns:x="adobe:ns:meta/">
     <rdf:RDF ...>
     <!-- RDF dokument uvedený výše -->
     </rdf:RDF>
   </x:xmpmeta>
<?xpacket end="w"?>
```

To mimojiné umožní snazší strojové vyhledávání a indexování zdrojů přítomných v LaTeXovém dokumentu.

Druhá ukázka použití

Mějme vytvořenu implementaci Dublin Core nového profilu užití metadat, například LOM. Autor LaTeXových balíků vytvoří balík, poskytující jednoduché příkazy propojující uživatelské rozhraní s abstraktním modelem. Příkladem nového elementu metadat by mohl být element LOM „Education.Difficulty“ [1], který by mohl být implementován jako jednoduchý příkaz `\difficulty`.

Třetí ukázka použití

V LaTeXu píšící autor vytváří nějaký dokument a přeje si, aby byl snadno citovatelný jinými autory. To se může v budoucnu bez problému stát požadavkem univerzit majících zvýšit jejich citační profil. Použije jednak běžné minimum:

```
\author{AN Other}
\title{All we know about Gnus}
```

Ale mimoto také `\includepackage{autocite}` (či jakkoli jinak bude balík pojmenován), čímž vygeneruje XMP metadata o autorovi, názvu, implicitním datu (`\today` ve formátu ISO 8601), jménu dokumentu (`\jobname`), rozsahu dokumentu (alespoň v bytech, ne-li ve slovech), typu dokumentu (kniha, článek atp.) a co možná nejvíce dalších informací, jež lze detekovat či usoudit, včetně metadat ze všech citací a hyperlinků na vnější zdroje.

Měly by být vytvořeny nádstavby nebo záplaty pro populární třídy dokumentů jako memoir [18], kluwer [15], elsevier [4] a jiných tak, aby bylo možné přidat metadata z jejich doplňujících polí, jako jsou `\submissiondate` a `\journalname`.

Seznam literatury [on-line 19. 5. 2009]

[1] Draft Standard for Learning Object Metadata. Autor: The Institute of Electrical and Electronics Engineers, Inc. [on-line]. URL: `http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf`

[2] Dublin Core Metadata Initiative (DCMI) [on-line].
URL: `http://dublincore.org/`

[3] Dublin Core Metadata Initiative Abstract Model. Tvůrci: Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnston a Thomas Baker [on-line].
URL: `http://dublincore.org/documents/abstract-model/`

[4] Elsevier – Main Homepage [on-line].
URL: `http://www.elsevier.com/`

[5] Expressing Dublin Core metadata using the DC-Text format [on-line].
Tvůrce: Pete Johnston.
URL: `http://dublincore.org/documents/dc-text/index.shtml`

[6] Expressing Dublin Core metadata using the Resource Description Framework. Tvůrci: Mikael Nilsson, Andy Powell, Pete Johnston a Ambjörn Naeve [on-line]. URL: `http://dublincore.org/documents/dc-rdf/`

[7] Extensible Metadata Platform (XMP). Tvůrce: Adobe Systems Incorporated [on-line]. URL: `http://www.adobe.com/products/xmp/`

[8] Google Summer of Code – TEX Users Group [on-line].
URL: `http://tug.org/gsoc/`

[9] Google Summer of Code discussions for TUG [on-line].
URL: `http://lists.tug.org/summer-of-code`

[10] Learning Object Metadata – Wikipedia [on-line].
URL: `http://en.wikipedia.org/wiki/Learning_object_metadata`

[11] Notation3 (N3) – A Readable RDF Syntax [on-line]. Editor: Tim Berners-Lee. URL: http://www.w3.org/DesignIssues/Notation3.html

[12] OpenURL ContextObject in SPAN (COinS): A Convention to Embed Bibliographic Metadata in HTML [on-line]. Napsal a editoval: Eric Hellman. URL: http://ocoins.info/

[13] Resource Description Framework (RDF) [on-line]. Tvůrci: Ivan Herman, Ralph Swick a Dan Brickley. URL: http://www.w3.org/RDF/

[14] RDF/XML Syntax Specification [on-line]. W3C Recommendation 10 February 2004. Editor: Dave Beckett. Editor celé série: Brian McBride. URL: http://www.w3.org/TR/rdf-syntax-grammar/

[15] SpringerLink – Main Homepage [on-line]. URL: http://www.springerlink.com/

[16] The hyperxmp package [on-line]. Tvůrce: Scott Pakin. URL: http://www.ctan.org/get/macros/latex/contrib/hyperxmp/

[17] The Joint DCMI/IEEE LTSC Taskforce – Wiki [on-line]. URL: http://dublincore.org/educationwiki/DCMIIEEELTSCTaskforce

[18] The memoir class for Configurable Typesetting [on-line]. User Guide. Autor: Peter Wilson. URL: http://www.ctan.org/tex-archive/macros/latex/contrib/memoir/memman.pdf

[19] XMP Specification [on-line]. Tvůrce: Adobe Systems Incorporated. URL: http://www.adobe.com/devnet/xmp/pdfs/xmp_specification.pdf

[20] Zotero – A Free Firefox Extension [on-line]. Vedoucí pracovníci projektu: Dan Cohen a Sean Takats. URL: http://www.zotero.org/

Summary: The Dublin Core Metadata Interface Project

The report brings a Czech translation of the web page describing one of the projects proposed for the GSoC 2009 with TUG as a mentoring organisation. The goal of the project was the implementation of the Dublin Core Abstract Model in TEX.

Key words: Google, Google Summer of Code 2009, TUG, TEX Users Group, Metadata, Dublin Core Abstract Model, RDF, XMP, XML.

Přeložil: Marcel Svitalský
marcel.svitalsky@centrum.cz