Zpravodaj Československého sdružení uživatelů TeXu

Marcel Svitalský
Google Summer of Code 2009 a TUG

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 19 (2009), No. 1-2, 94–101

Persistent URL: http://dml.cz/dmlcz/150036

Google Summer of Code 2009 a TUG

Marcel Svitalský

Abstrakt

Zpráva přináší překlad webové stránky s idejemi TUGu [9] pro Google Summer of Code 2009 [7].

Klíčová slova: Google, Google Summer of Code 2009, TUG, TEX Users Group.

Ročník 2009

TUG se letos nezúčastní programu Google Summer of Code [7] v roli poradní organizace [16]; naše žádost byla pro letošní rok zamítnuta, avšak kterýkoli zájemce si může vyhledat seznam ideí, jejž jsme společně vytvořili, a nechat se jím inspirovat. V TEXovém světě zbývá stále ještě víc než dost vývojářské práce, již je třeba udělat.

V roce 2008 se pod záštitou TUGu zúčastnili GSoC tři studenti. Jejich projekty [18] a kódy [19], které vytvořili, stejně jako oznámení TUGu pro rok 2008 [8], si můžete v angličtině pročíst a prohlédnout na Internetu.

Všechny organizační diskuse TUGu týkající se GSoC se odehrávají na mailing listu `summer-of-code@tug.org` [10]; bez obav se do něj můžete přihlásit nebo jeho archivy [21] pročítat.

Nápady a myšlenky projektů byly:

1. Přístupné PDF s pomocí TEXu

Cílem tohoto projektu je obohacení pdfTEXu o schopnost produkovat označkované PDF ve shodě se specifikacemi PDF/A, PDF/UA, ISO 32000 a (dosud nepublikovanou) ISO-32000-2. To zahrnuje jak strukturní, tak obsahové značkování pro užití se čtečkami obrazovky (pro zrakově postižené) a softwarovými pluginy schopnými zobrazit a vylepšit strukturu i obsah matematických vzorců.

Úspěšné završení tohoto projektu musí zahrnovat následující úlohy:

a) Vyhledejte všechna místa ve standardním LaTeXovém kódu, kde je tagování struktury nebo obsahu vhodné.
b) Vytvořte soubor(y) obsahující alternativní definice maker zahrnující programové „přípojky“ umožňující umístění tagů na patřičná místa ve výstupním streamu, tj. do TeXového „vertikálního seznamu“.
c) Vyhledejte ta místa, kde upravená makra nebudou postačující ke vložení strukturních nebo obsahových tagů optimálním způsobem. To bude indikovat potřebu přidat pod nimi běžícímu procesoru dodatečné schopnosti.
d) Dodejte soubor definic (La)TeXových maker, poskytujících všechny strukturní, obsahové a matematické tagy, podporované již ve specifikaci PDF 1.7 (nebo vyšší) a jež budou podporovány normou ISO-32000-2.
e) Dodejte „podpůrné“ soubory umožňující propojení „přípojek“ z bodu b) s definicemi maker z bodu d). Takových podpůrných souborů může být zapotřebí více, aby poskytovaly různé úrovně podpory vyhovující požadavkům různých úrovní tagování, kupř. pro PDF/X, PDF/A (různých verzí), PDF/UA a pro plné tagování matematických vzorců.
f) Dodejte soubory/ovladače, řídící způsob, jímž „přípojky“ z bodu e) vkládají tagy do výstupu vytvářeného určitým procesorem. Zpočátku bude třeba podporovat pouze jeden procesor, např. pdfTeX, avšak kód by měl být natolik modulární, aby podporu dalších procesorů, např. XeTeXu a LuaTeXu, bylo možné zajistit pouhou výměnou souboru/ovladače. Všechny další moduly by mělo být možné užít bez jakýchkoli úprav.
g) Vytvořte patřičnou dokumentaci popisující, co mají jednotlivé moduly dělat a jak jich užívat.

Naplnění těchto zadání bude společným úsilím zahrnujícím všechny účastníky projektu, vizte níže, nikoliv pouze studenta. Student by měl nepochybně být dobře obeznámen s postupy programování TeXových maker, jakož i se způsoby, jimiž jsou tato užívána ve vnitřních makrech LaTeXu.

Znalost dalších aspektů programování, jako PDF/PostScript, Tangle & Weave či programování pro pluginy Adobe, bude považována za bonus, který se může ukázat užitečným, není však nezbytným. Diskuse k této problematice hostí TeXový mailing list [11] v River Valley Technologies.

Personální obsazení (poradci):

* Hàn Thế Thành, River Valley Technologies, Německo. Vývojář pdfTeXu.
* Ross Moore, Senior Lecturer in Mathematics, Macquarie University, Sydney, Australia. TeXové/LaTeXové programování, předsednictvo TUGu.
* Neil Soiffer, Senior Scientist, Design Science Inc., California and St. Paul, Minnesota, USA. Plug-in software k podpoře matematického obsahu.
* CV Radhakrishnan, River Valley Technologies, India, Trivandrum, Kerala, India. TeXové/LaTeXové programování, publikování vědeckých prací.

2. Dublin Core metadata a TEX

Celý text návrhu [5] je dostupný na Internetu a v samostatném překladu v tomto Zpravodaji, zde uvádíme shrnutí. S projektem Dublin Core se lze v češtině podrobně seznámit na stránkách [4].

Dublin Core Metadata Initiative je otevřená organizace zapojená do vývoje standardů pro interoperabilní online metadata podporujících široké rozpětí účelů a aplikačních modelů. Vyvinula abstraktní framework pro metadata a několik strojově čitelných reprezentací metadatových výrazů, mimo jiné i ty v Resource Description Frameworku (RDF).

Jedním z významných uživatelů RDF metadat je Adobe, tvůrce formátu PDF. Jejich eXtensible Metadata Platform (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.

Klíčovými výstupy tohoto projektu by měly být:

1. TEXová implementace Dublin Core Abstract modelu;
2. metody pro export metadat z abstraktního modelu do vnějších souborů v různých formátech, zejména RDF+XML, popř. také DC-TEXT a N3;
3. automatické vkládání XMP paketů do vytvářeného PDF souboru v případě pdfLATEXu, s defaultním minimem XMP výrazů Z39.88 OpenURL COinS polí, jak pro vlastní metadata dokumentu, tak i pro všechny citované odkazy a vnější 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 příslušné oblasti jejich zájmu. Osobně uvažuji o Learning Object Metadata, avšak mapování LOM do Dublin Core je problematické.

Poradci projektu budou Peter Flynn a Matthew Leingang.

3. Rozpoznání rukou psaných LATEXových symbolů

Tento projekt dosud nemá přiřazeného poradce. Kdokoliv, kdo by měl zájem jako poradce působit, je srdečně zván ke kontaktu prostřednictvím mailing listu [10].

Sázecí systém LATEX [15] poskytuje příkazy pro sazbu tisíců rozličných symbolů potřebných pro přípravu dokumentů z oblastí lingvistiky, matematiky, hudby, techniky, fyziky a mnoha dalších. Najít LATEXový název pro daný glyf může být

pro tvůrce dokumentu obtížným úkolem. V současnosti je nejlepším řešením užít Comprehensive LaTeX Symbol List [3] (úplný seznam LaTeXových symbolů), souhrn tabulek symbolů organizovaných do ad hoc kategorií a indexovaných dle LaTeXového názvu symbolu. Problémem tohoto přístupu je, že různí uživatelé spojují s týmž glyfem odlišné názvy, čímž se vyhledávání stává obtížným.

Uvažte pro příklad, že se pokoušíte vyhledat LaTeXový název pro kroužek s tečkou uprostřed. Astronom bude možná hledat „slunce“, matematik „mez“, lingvista „mlaskavou souhlásku“, tvůrce map „městské centrum“, někdo píšící o alchymii „zlato“. Ve skutečnosti je výpisu významů tohoto symbolu věnována celá stránka Wikipedie [1]. Neanglicky mluvící budou dále znevýhodněni tím, že většina LaTeXových symbolů byla pojmenována anglicky.

Domníváme se, že uživatelům LaTeXu by byl velkou nápomocí webový nástroj vyhledávající symboly založený na rozpoznávání textu. To jest, představujeme si webovou stránku, na níž by uživatel mohl nakreslit svůj symbol a následně by mu byl zobrazen seznam LaTeXových symbolů (příkazů a jejich výstupů) nejvíce odpovídajících uživatelově kresbě.

Student by musel vyhodnotit četné možnosti při rozpoznávání ručně vykreslených symbolů, najít vhodnou vnitřní representaci pro tisíce LaTeXových symbolů a asociovaná metadata, vytvořit vhodné uživatelské rozhraní pro práci s rozpoznávačem symbolů a zajistit, že výsledný software bude spravovatelný, zejména s přihlédnutím k četnosti přidávání nových symbolů do LaTeXu.

Taková idea je nepochybně na implementaci náročná. Zároveň to však bude zcela jistě vzrušující a prospěšná zkušenost, vzhledem k přehršli technologií, jichž se bude týkat: TeX/LaTeX, rozpoznávání textu, rozličné webové technologie, a patrně řada programovacích jazyků. Lze se zde mnoho naučit a TUG velmi rád bude pomáhat studentovi se zájmem a schopností uskutečnit projekt převodu rukopisných symbolů do LaTeXu.

Návrh podal Scott Pakin, je však třeba poradce.

4. Zvýraznění syntaxe s hyperlinkem pro TeX

Nyní je běžné, že zobrazuje-li se kód na webové stránce, je užito zvýraznění syntaxe. Děje se tak např. na stránce Google Code [6].

Tento projekt má poskytnout zvýraznění syntaxe pro TeXový kód – jak pro dokumenty, tak pro makra – s další přidanou schopností. Každý zvýrazněný příkaz bude zároveň hyperlinkem nabízejícím bublinovou nápovědu a, pokud naň bude kliknuto, vedoucím k další dokumentaci.

Dvěma z nejlepších zvýrazňovačů syntaxe jsou:

- Pygments [20] v jazyku Python, a
- Chili [17] pro JavaScript, založeném na jQuery.

Tento projekt sestává ze tří částí. Prvou je poskytnutí pokročilého zvýrazňování syntaxe pro TEXový kód. Druhou je vytvoření databáze příkazů. Třetí je vzájemné spojení prvé a druhé části.

V závislosti na obtížích, jež se vyskytnou, se tento projekt může ukázat jako příliš rozsáhlý. Pokud by to nastalo, předpokládáme, že student uskuteční pouze jeho část.

Poradcem projektu bude Jonathan Fine.

5. Nové šablony dokumentů pro LaTeX

Lamport vytvořil pro LaTeX řadu šablon dokumentů, jako jsou třídy book, article atd. Dokonce i dnes užívá velké množství LaTeXových dokumentů tyto šablony, což ústí ve zřetelný „LaTeXový vzhled“.

Autorům jsou rovněž k disposici třídy pro specializovaná užití, např. pro časopisy, konkrétní konference nebo diplomové práce té či oné univerzity. Avšak tyto třídy jsou velmi často jen uzpůsobenými Lamportovými šablonami a udržují si svůj vzhled. Z opačné strany vzato by bylo vhodnou součástí tohoto projektu prozkoumání existujících tříd poskytujících výrazně odlišný vzhled, jako jsou koma-scripty scr* [14] či smfart [2] Francouzské matematické společnosti.

Tento projekt by měl poskytnout alternativní šablony pro široké užití. Tyto šablony mohou být vhodné i pro knihy a články, popř. pro další účely. Měly by pokud možno být doprovázeny návodem pro možné uživatele s ohledem na jejich vhodné použití, např. pro knihy převážně bez matematiky či pro automaticky generované texty.

Poradcem projektu bude Jim Hefferon.

6. Implementace fontspec pro LuaTeX

XeTeX je populárním rozšířením TEXu umožňujícím užití TrueType a OpenType písem, aniž by musely být upraveny do Type 1 nebo TFM souborů. Za část svého úspěchu vděčí balíku fontspec Willa Robertsona, poskytující LaTeXové rozhraní primitivům XeTeXu pro nahrání písem. Cílem tohoto projektu je implementovat balík fontspec pro LuaTeX.

Jelikož XeTeX a LuaTeX představují velmi odlišná paradigmata v přístupu k rozšiřování TEXu, LuaTeXová část fontspec bude výrazně odlišná od nyní existujícího kódu pro XeTeX: zatímco tento využívá řady dodatečných systémových knihoven, aby TEXu dodal pokročilou podporu písem, LuaTeX poskytuje spojení mezi TEXem a jazykem Lua a umožňuje zapojit kód Lua do TEXového sázecího procesoru. Tudíž převážná většina práce bude spočívat v implementaci pokročilých rysů písem v Lua. Tedy toho, co již dělá ConTeXt verze „Mark IV“.

Je třeba věnovat pozornost těmto aspektům:

Vyhledávání písem: Aby LuaTEX dokázal užívat systémová písma stejně jako XeTEX, je třeba pro ně implementovat vyhledávací mechanismus. Nabízí se několik možností:

- XeTEX je slinkován s knihovnou fontconfig, což ale není případem LuaTEXu; avšak na systémech, kde je fontconfig k dispozici také jako samostatný program, jej lze spustit a použít jeho výstup.
- Mohlo by být možné přeložit fontconfig jako sdílenou knihovnu a tu užívat z LuaTEXu. Lua totiž disponuje vydařeným rozhraním pro jazyk C umožňujícím mu přímo užívat knihovny tohoto jazyka. Zřejmým problémem takového přístupu by však byly značné potíže s přenositelností mezi různými operačními systémy.
- Bylo by také možné implementovat vlastní jednoduchý vyhledávací mechanismus, prohledávající adresáře standardní pro jednotlivé platformy, konkrétně C:\Windows\Fonts pod Microsoft Windows, /System/Library/Fonts, /Library/Fonts a ∼/Library/Fonts v Mac OS, /usr/share/fonts v Linuxu atd. Takto to dělá ConTEXt Mark IV: spouští skript kešující všechny soubory a jména písem.

Tyto schopnosti by samozřejmě byly doplněním knihovny kpathsea, kterou lze také volat z Lua. Měli bychom tudíž duální situaci velmi podobnou XeTEXu.

Nahrávání písma: LuaTEXový primitiv `\font` se na úrovni procesoru chová přibližně stejně jako týž primitiv v pdfTEXu, a nemůže tedy nahrát soubory TrueType nebo OpenType písem. Očekává defaultní typ souboru TFM. Potřebujeme proto „přetížit“ tento primitiv tak, aby emuloval chování stejné jako v XeTEXu.

Rozložení OpenType: I zde je několik možností:

- LuaTEX disponuje nástroji k implementaci kompletního procesoru rozložení OpenType. Vyžadovalo by to značné úsilí, avšak základ kódu je již nyní dostupný pro ConTEXt a bylo by možné ho s výhodou využít.
- XeTEX používá externí knihovny jako je International Components for Unicode (ICU) [12] od IBM. Podobně jako ve druhé možnosti prvého bodu bychom mohli zkompilovat sdílenou knihovnu obsahující wrapper v jazyce C poskytující rozhraní pro Lua.

LaTEXové rozhraní: V ideálním případě by LuaTEXová implementace fontspec měla mít stejné vysokoúrovňové rozhraní jako současný balík.

Poradci projektu budou Will Robertson a Arthur Reutenauer.

7. Výchozí mikrokernel LaTeX3

Sázecí systém LaTeX po mnoho let znamenal LaTeX 2$_\varepsilon$. Poslední vývoj následníka, LaTeX3 [13], byl soustředěn převážně na nový nízkoúrovňový programovací systém pro TeX. Jelikož tato nízkoúrovňová práce již dozrává, lze zkoušet nové programovací postupy pro práci na vyšší úrovni.

Cílem projektu mikrokernelu LaTeX3 je započít zkoumání, jak lze uplatnit nízkoúrovňový systém k vytvoření systému, který bude možno použít k vysázení jednoduchého dokumentu ve stylu LaTeX 2$_\varepsilon$, aniž by bylo zapotřebí nahrát navrch současný LaTeX 2$_\varepsilon$ kernel. Jako úvodní cíl bude použit k otestování základní dokument:

```
\documentclass{minimal}
\begin{document}
\emph{Hello World!}
\end{document}
```

Zde popsaný mikrokernel nemá být úplnou implementací kernelu LaTeX 2$_\varepsilon$, tj. `latex.ltx`. Existuje řada dosud nezodpovězených otázek ohledně uživatelského rozhraní LaTeX3. Cílem projektu je vytvořit základní samostatný kernel, jenž poté bude možno pomalu rozšiřovat implementováním dalších vlastností. Nejpravděpodobněji z `latex.ltx`, možná i z přídavných LaTeXových balíků.

Dalším logickým krokem po vytvoření systému schopného pracovat s testovacím dokumentem budou dodatky jako základní strukturní příkazy a prostředí (seznamy, zarovnání atp.). Jisté oblasti mohou být také vyloučeny: New Font Selection Scheme, komplexní výstupní rutiny či plovoucí obsah jsou všechny mimo rámec tohoto projektu.

Poradcem projektu bude Joseph Wright.

Seznam literatury [on-line 14. 5. 2009]

[1] Circled dot – Wikipedia, the free encyclopedia [on-line]. URL: `http://en.wikipedia.org/wiki/Circle_with_a_point_at_its_centre`

[2] Classes for Société mathématique de France publications: smflatex package. `http://www.ctan.org/tex-archive/macros/latex/contrib/smflatex/`

[3] CTAN: The Comprehensive LaTeX Symbol List [on-line]. URL: `http://www.ctan.org/tex-archive/info/symbols/comprehensive/`

[4] Dublin Core – Czech homepage [on-line]. URL: `http://www.ics.muni.cz/dublin_core/`

[5] Dublin Core metadata project [on-line]. URL: `http://tug.org/gsoc/dublincore.html`

[6] Google Code [on-line]. URL: `http://code.google.com/`

[7] Google Open Source Programs [on-line].
URL: http://socghop.appspot.com/
[8] Google Summer of Code 2008 and TUG [on-line].
URL: http://tug.org/gsoc/2008.html
[9] Google Summer of Code and TUG [on-line].
URL: http://tug.org/gsoc/
[10] Google Summer of Code discussions for TUG [on-line].
URL: http://lists.tug.org/summer-of-code
[11] TEX Info Page [on-line]. URL:
http://lists.river-valley.com/cgi-bin/mailman/listinfo/tex
[12] International Components for Unicode [on-line]. Home Page.
URL: http://www.icu-project.org/
[13] LATEX project: The LATEX3 project [on-line].
URL: http://www.latex-project.org/latex3.html
[14] KOMA-Script: A versatile bundle of LATEX document classes and packages.
http://ctan.org/tex-archive/macros/latex/contrib/koma-script/
[15] LATEX project: LATEX – A document preparation system [on-line].
URL: http://www.latex-project.org/
[16] List participating Organizations [on-line]. URL: http://socghop.
appspot.com/program/accepted_orgs/google/gsoc2009
[17] Notes Log >> Chili [on-line].
URL: http://noteslog.com/category/chili/
[18] Organization Information: TEX Users Group [on-line].
URL: http://code.google.com/soc/2008/tex/about.html
[19] Project Home: Code samples from students working with TEX Users Group
for Google Summer of Code 2008 [on-line].
URL: http://code.google.com/p/google-summer-of-code-2008-tex/
[20] Pygments – Python syntax highlighter [on-line].
URL: http://pygments.org/
[21] The summer-of-code Archives [on-line].
URL: http://tug.org/pipermail/summer-of-code/

Summary: Google Summer of Code 2009 and TUG

The report is a Czech translation of the web page presenting ideas of the TEX Users Group for Google Summer of Code 2009.

Key words: Google, Google Summer of Code 2009, TUG, TEX Users Group.

Přeložil: Marcel Svitalský
marcel.svitalsky@centrum.cz