Zpravodaj Československého sdružení uživatelů TeXu

Petr Sojka; Michal Růžička

# Redakční systém odborného časopisu s podporou exportu do digitální knihovny

*Zpravodaj Československého sdružení uživatelů TeXu*, Vol. 21 (2011), No. 1, 4–20

Persistent URL: http://dml.cz/dmlcz/150152

Budeme vděčni i za pomocnou ruku v získávání nových členů, propagaci sdružení, editaci a redakci tematických webových stránek na připravovaném novém webu sdružení, ale i za konstriktivní náměty na nové aktivity či připomínky ke stávajícímu chodu sdružení.

*Masarykova univerzita, Fakulta informatiky, Botanická 68a, 602 00 Brno*
*<sojka@fi.muni.cz>*

# Redakční systém odborného časopisu s podporou exportu do digitální knihovny

Petr Sojka, Michal Růžička

Postup zpracování odborných, zejména matematických časopisů je založen na TEXu a navazujících technologiích. Vydavatel většinou články zpřístupňuje a paralelně publikuje i elektronicky. Vytváří optimalizovanou verzi pro webové šíření, verzi pro archivaci, případně i verzi optimalizovanou pro čtení na obrazovce.

Článek popisuje navržený a implementovaný postup zpracování několika matematických časopisů deponujících svou produkci v České digitální matematické knihovně DML-CZ a následně v Evropské digitální matematické knihovně EuDML.

## Publikování odborných časopisů

V akademickém světě se z mnoha stran ozývá známé „publikuj nebo zmiz“ (“Publish or Perish”). Množství každoročně publikovaných článků roste exponenciálně, a oblast technických věd a matematiky není výjimkou. Je třeba publikovat rychleji a kvalitně poskytovat metadata pro vyhledávací stroje, neboť většina uživatelů k článkům přistupuje přes vyhledávací portály. Velké procento odborných publikací a časopisů, zejména matematických, je sázeno TEXem. V Indii, Litvě či jinde po světě jsou firmy se stovkami zaměstnanců, které tyto publikace sázejí a zpracovávají pro vydavatele jako Elsevier nebo Springer.

Vydavatelé jednak svou produkci archivují v systémech jako Portico[1] pro případy katastrofických scénářů, jednak zpřístupňují předplatitelům a Google ve svých digitálních knihovnách jako je SpringerLink[2] nebo ACM Digital Library.

[1] <http://www.portico.org/>
[2] <http://www.springerlink.com/>

Nadpoloviční většina čtenářů své odborné dokumenty najde přes Google či Google Scholar, takže je zásadní, aby plné texty byly stahovačům Google známy a poskytnuty, i když jsou za předplatitelskou posuvnou zdí, anglicky *moving wall* (princip, kdy starší články jsou dostupné volně, ale novějších několik ročníku je dostupných jen z rozsahů internetových adres předplatitelů).

V Česku a Slovensku je publikována více než desítka odborných matematických časopisů, a všechny redakce k sazbě používají TEX. V roce 2005 se podařilo získat grant na projekt České digitální matematické knihovny DML-CZ (`<http://dml.cz>`) [1], která by vydávané články zpřístupňovala odborné komunitě a vystavovala pro indexování Google (Scholar). Projekt byl úspěšný, a v roce 2010 na něho již navazuje projekt Evropské digitální matematické knihovny EuDML (`<http://eudml.eu>`), kterým se přes 300 000 stran odborných matematických textů dále šíří ke čtenářům.

Cílem projektu DML-CZ bylo zpracovat jednak články digitalizací (retro-digital období), ale také články u kterých již byla, ale často neúplná, primární data (retro-born-digital období), a přebírat co nejvíce automatizovaně od redakcí časopisů data nově vydávaných čísel (tzv. *born-digital*, vznikajících elektronicky).

## Born-digital systém DML-CZ

V rámci projektu DML-CZ bylo navržen způsob zpracování budoucích čísel časopisů a jejich zařazování do repositáře projektu. Jelikož jsou matematické časopisy typicky sázeny TEXem, jsou i born-digital systémy projektu DML-CZ do značné míry postaveny na TEXových technologiích.

Hlavní myšlenkou bylo získávání born-digital dat pro DML-CZ jako vedlejšího produktu práce redakcí při vydávání nových čísel svých časopisů. Zde bylo možné dát se dvěma cestami – vytvořit nový komplexní redakční systém, který se postará nejen o vygenerování DML-CZ metadat, ale také automatizuje co nejvíce činností prováděných při přípravě nového čísla matematického časopisu, anebo provést jen minimální zásahy do zaběhnutého redakčního workflow a s pokud možno minimálními nároky na redakci jej obohatit o generování výstupů pro projekt DML-CZ.

### Komplexní born-digital systém

Pilotním projektem přechodu do born-digital režimu byl brněnský matematický časopis Archivum Mathematicum vydávaný na Přírodovědecké fakultě Masarykovy univerzity. Zde jsme se vydali první cestou – připravili jsme zbrusu nový komplexní systém. Inspirací nám byl systém [2] používaným ve francouzském projektu CEDRAM [3].

V tomto systému redaktor všechny články upraví do požadovaného tvaru za použití značkování, které se příliš neliší od standardního značkování LaTeXu a společně s případnými externími soubory (obrázky apod.) je umístí do odpovídající adresářové struktury. Pro každé číslo následně připraví jednoduchý TeXový dokument – řídící soubor, ve kterém definuje pořadí jednotlivých článků, ročník, číslo a rok vydání daného čísla časopisu apod. V tomto bodě je již redakční systém schopen automaticky

- přeložit všechny články čísla a sesadit je do jediného výsledného dokumentu ve formátech PDF a PostScript,
- vygenerovat obálku čísla s obsahem a korektními údaji o ročníku, čísle atd. ve formátech PDF a PostScript,
- vygenerovat zrcadlové tiskové předlohy s ořezovými značkami ve formátech PDF a PostScript,
- vygenerovat elektronickou podobu časopisu pro vystavení na webu,
- vygenerovat formuláře pro recenzní řízení,
- vygenerovat (a na požádání automaticky odeslat) metadata pro projekt DML-CZ,
- vygenerovat další pomocné výstupy.

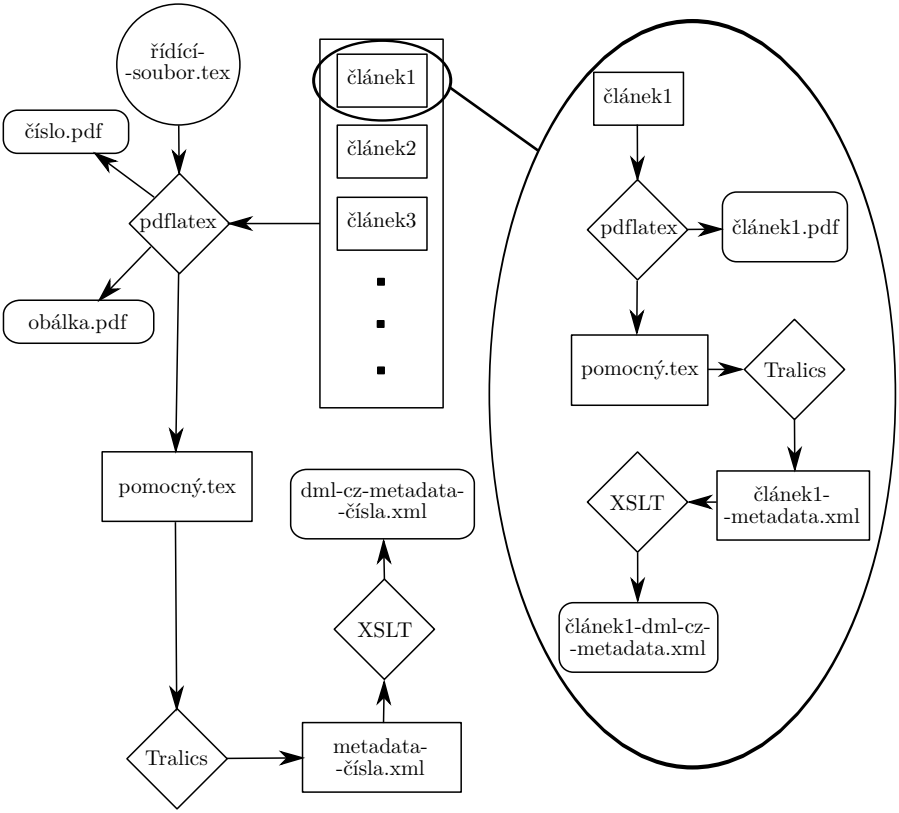
Schéma práce born-digital redakčního systému DML-CZ můžete vidět na obrázku 1.

Celé zpracování je řízeno přímo ze třídy dokumentu pomocí TeXového příkazu `\write18`[3], ke spuštění celého automatizovaného zpracování proto uživateli stačí přeložit TeXem řídící soubor čísla. Překlad jednotlivých článků probíhá zcela odděleně (systém řeší nastavení správného čísla první stran každého článku apod.), nehrozí proto konflikty mezi makry používanými v jednotlivých článcích. Výsledné číslo časopisu je vytvořeno spojením jednotlivých výstupních PDF dokumentů.

Jádro systému generující DML-CZ metadata ve formátu XML je postaveno na open-source nástroji Tralics [4, 5] (jehož použití je blíže popsáno v dalším oddílu tohoto článku na straně 8). Tralics zpracovává jen soubory speciálně pro něj připravené – jedná se o LaTeXové dokumenty s minimálním dodatečným značkováním, které vznikly při překladu článků díky speciální definici některých v článcích používaných maker. Tyto minimální dokumenty obsahují původní LaTeXový zdrojový text položek, které se mají stát součástí metadat. Pro zjednodušení konfigurace Tralicsu nejsou produkována metadata přímo v konečném DML-CZ značkování. Tralicsem vyprodukovaný XML dokument je do finální podoby převeden pomocí XLST.

DML-CZ metadata mají dvě části – popis vlastního článku (název, autoři, jazyk článku, abstrakt atd.) a seznam v článku citované literatury. Pro použití

[3]Tento příkaz umožňuje uživateli vykonávat běžné příkazy operačního systému přímo z TeXového zdrojového textu. Z bezpečnostních důvodů musí být použití uživatelem explicitně povoleno přidáním `-shell-escape` nebo obdobného argumentu na příkazovém řádku při spuštění TeXu.




Obrázek 1: Schéma práce DML-CZ born-digital systému

v projektu DML-CZ (propojení všech článků stejných autorů, obsahová klasifikace článků, vyhledávání dle kritérií atd.) je důležité, aby byla metadata článků kvalitně strukturně označena. Získání strukturně označkovaných metadat není příliš obtížné, neboť logické oddělení jednotlivých elementů (`\title`, `\author` atd.) je do značné míry podporováno přímo standardními třídami dokumentů LaTeXu. Horší situace panuje v oblasti seznamů citované literatury, kde LaTeXové prostředí `thebibliography` žádnou strukturu jednotlivých záznamů nevyžaduje.

Poměrně rozšířeným způsobem zpracování bibliografických záznamů je však program BibTeX, jehož bibliografické databáze jsou navíc přímo podporovány v programu Tralics. V případě komplexního born-digital redakčního systému bylo proto přikročeno k použití právě tohoto nástroje. Výhodou je dodržení jednotné

podoby seznamů literatury u všech článků, a zároveň jsou automaticky připravena kvalitně strukturně označená data pro generování DML-CZ metadat.

**Minimalistický born-digital systém**
Některé redakce již dlouhou dobu používají svůj osvědčený zaběhnutý způsob práce, a tak i když se chtěly zapojit jako born-digital přispěvatelé do projektu DML-CZ, místo zavedení nového integrovaného systému dávaly přednost druhé cestě generování DML-CZ podkladů – zavedení minimálního rozšíření stávajícího workflow. Právě zde jsme opět použili osvědčenou komponentu integrovaného systému, nástroj Tralics – ukázka jeho použití je obsahem následujícího oddílu.

Stejně jako v případě komplexního born-digital systému bylo nutné drobně modifikovat používané značkování s ohledem na strukturní vyznačení jednotlivých datových položek pro DML-CZ metadata, pro sazbu seznamů literatury se však již nepoužívá BibTEX, ale bylo zavedeno speciální strukturní značení.

Ať se již redakce rozhodla pro kteroukoliv cestu, pro projekt DML-CZ je velmi výhodné, že data vznikají přímo při přípravě nového čísla z původních zdrojových textů – je tak zaručeno, že se v digitální knihovně objeví články ve stejné podobě, ve které byly vytištěny.

## Tralics

Tralics (`<http://www-sop.inria.fr/apics/tralics/>`) je konvertor LaTeXu do XML. Jedná se o původem francouzský open-source nástroj šířený pod CeCILL licencí.[4]

Tralics je multiplatformní, kromě zdrojových textů jsou k dipozici i předpřipravené zkompilované binární verze tohoto nástroje pro operační systém GNU/Linux, Apple Mac OS X a Microsoft Windows.[5]

Na rozdíl od některých jiných nástrojů, např. TeX4ht [6], je Tralics primárně určen ke konverzi dokumentů do obecného strukturně označkovaného XML, nikoliv ke konverzi TeXových dokumentů do jejich vizuálně obdobné podoby v jiném formátu (např. HTML). Také pracuje odlišným způsobem.

TeX4ht kupříkladu vstupní dokument překládá přímo TeXem do DVI. Načtena je pouze speciální sada maker, která do výstupního souboru přidá `\special{...}` příkazy s poznámkami pro TeX4ht. Tyto poznámky jsou při postprocessingu takto vytvořeného DVI souboru využity k vytvoření HTML podoby dokumentu. [7]

Tralics pracuje odlišným způsobem – Tralics sám překládá přímo zdrojový text LaTeXového dokumentu. Je přitom schopen provádět expanzi i složitých uživatelem

[4] CeCILL licence (`<http://www.cecill.info/>`) je podobná a kompatibilní s GNU GPL, zohledňuje však specifika francouzského právního systému.

[5] Pro běh na MS Windows je použito knihoven z projektu Cygwin (`<http://www.cygwin.com/>`).

definovaných maker apod. Mimoto dokáže zpracovávat i bibliografické databáze BibTEXu. Pokud je seznam literatury do dokumentu vkládán tímto způsobem, je Tralicsem zpracován v jediném kroku zároveň s dokumentem samotným.

Schopnost přímého zpracování LATEXového zdrojového textu a TEXové makroexpanze je velmi významným argumentem pro použití Tralicsu, neboť díky tomu podává velmi dobré výsledky a je možná podrobná konfigurace konverze. Právě mnohotvárnost zápisu i jednoduchých konstrukcí v TEXu je důvodem, proč by bylo obtížené získávat data (byť i jejich malou podmnožinu potřebnou pro DML-CZ) přímo ze zdrojových textů jiným způsobem, např. skriptem v Perlu. Použití Tralicsu je mnohem pohodlnější a výsledky lepší.

**Konfigurace Tralicsu**

Konfigurace Tralicsu je pro uživatele TEXu velmi přirozená. Pro makra původního dokumentu definuje jejich transformaci do XML standardním TEXovým zápisem. Tralics k tomuto účelu poskytuje sadu nových TEXových příkazů, kterými je možné definovat podobu XML výstupu.

V konfiguračních souborech jsou definice maker uloženy v souborech kopírujících strukturu jejich standardních LATEXových protějšků – LATEXové třídě dokumentu / balíčku (`.cls`/`.sty`) odpovídá `.clt`/`.plt` konfigurační soubor Tralicsu se stejným jménem. Možná je samozřejmě definice i bez vazby na nějakou třídu dokumentu / balík maker. Součástí balíku se zdrojovými texty Tralicsu je také sada konfiguračních souborů pro základní třídy dokumentů a makrobalíky LATEXu.

**Ukázka použití Tralicsu**

Ačkoliv jsou na domovské stránce Tralicsu (`<http://www-sop.inria.fr/apics/tralics/>`) k dispozici předkompilované spustitelné soubory (obzvláště užitečné na platformě MS Windows, kde standardně není k dispozici překladač C++[6]), je vždy vhodné stáhnout také archiv se zdrojovými texty dané verze programu. Součástí archivu (adresář `confdir/`) je totiž také sada předpřipravených konfiguračních souborů zmíněných výše a sada testů (spouštěná skriptem `Test/alltests`) pro ověření funkčnosti Tralicsu.

Kompilace ze zdrojových textů v unixovém operačním systému je jednoduchá a vyžaduje jen překladač jazyka C++. Po rozbalení archivu stačí přejít do adresáře `src/` a příkazem `make` spustit kompilaci. Výsledkem je spustitelný soubor

[6]Na platformě Windows bude ke spuštění programu potřeba také dynamická knihovna `cygwin1.dll`. K provozování Tralicsu není třeba instalovat celé prostředí Cygwin (`<http://www.cygwin.com/>`), potřebnou knihovnu stačí umístit do stejného adresáře jako zkompilovaný Tralics. Samotná knihovna je součástí balíku `cygwin-<verze>.tar.bz2`, který je ke stažení na některém zrcadle FTP serveru projektu (`<http://www.cygwin.com/mirrors.html>`), v České republice např. na adrese `<ftp://ftp.fit.vutbr.cz/pub/systems/cygwin/release/cygwin/>`.

`src/tralics`. Zkompilovaný program můžeme v tomto umístění rovnou otestovat, pokud přejdeme do adresáře `Test` a spustíme skript `alltests`. Skript programem `src/tralics` přeloží TEXové soubory v adresáři `Test` a XML výstupy porovná se vzorovými soubory v adresáři `Modele`. Pokud testy proběhnou bez chyb, máme Tralics připraven k použití – potřebovat budeme zkompilovaný program `tralics` adresář se standardní konfigurací `confdir/`, které si můžeme přesunout dle svého uvážení kamkoliv do systému.

```
~/tralics-2.13.6$ mkdir ~/tralics
~/tralics-2.13.6$ mv confdir/ src/tralics ~/tralics/
~/tralics-2.13.6$ cd ~/tralics/
```

Nyní si můžeme vyzkoušet konverzi jednoduchého LATEXového dokumentu do XML.

```
~/tralics$ vim helloworld.tex
            ...
~/tralics$ cat helloworld.tex
\documentclass{article}
\def\hello{\uppercase{h}ello}
\begin{document}
\hello{} world!

Příliš žluťoučký kůň úpěl ďábelské ódy.
\end{document}
```

Dokument můžeme přeložit příkazem:

```
~/tralics$ ./tralics -confdir=confdir/ -utf8 -utf8output helloworld.tex
```

Parametrem `-confdir` přímo při spuštění programu určujeme umístění adresáře s konfiguračními soubory, nezáleží proto na jeho umístění v systému. Použity jsou také argumenty pro volbu vstupního a výstupního kódování (`-utf8`, `-utf8output`). Tralics podporuje jen omezenou množinu kódování, je však mezi nimi také UTF-8. Nejjednodušším postupem proto je případně zkonvertovat vstupní dokumenty Tralicsu právě do tohoto kódování. Při znalosti původního kódování dokumentu je toto možné snadno provést automatizovaně, UTF-8 navíc dokáže kódovat libovolný znak Unicode, problém proto nebude činit žádné vstupní kódování původního dokumentu.

Výsledkem překladu je soubor `helloworld.xml`:

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<std>
<p>Hello world!</p>
<p>Příliš žluťoučký kůň úpěl ďábelské ódy.
</p>
</std>
```

Jedná se tedy skutečně o strukturně označkovaný XML přepis původního dokumentu. Je také vidět, že Tralics správně zpracoval uživatelem definované makro `\hello`.

Deklarace typu dokumentu odkazuje na DTD `classes.dtd`, který však není součástí standardní konfigurace. Gramatiku našich dokumentu bychom museli definovat sami. Patřičným nastavením (např. v konfiguračním souboru) lze odkazované DTD snadno změnit za jiné. Tamtéž lze definovat také příkazy nezávisle na konkrétní třídě dokumentu / balíku maker.

```
~/tralics$ vim hwconfig.tcf
            ...
~/tralics$ cat hwconfig.tcf
DocType = hello world.dtd

BeginAlias
  std report book article minimal
End

BeginCommands
  \def\world{world}
End

~/tralics$ vim helloworld.tex
            ...
~/tralics$ cat helloworld.tex
\documentclass{article}
\def\hello{\uppercase{h}ello}
\begin{document}
\hello{} \world!

Příliš žluťoučký kůň úpěl ďábelské ódy.
\end{document}

~/tralics$ ./tralics -confdir=confdir/ -configfile=./hwconfig.tcf \
> -utf8 -utf8output helloworld.tex
            ...
Output written on helloworld.xml (234 bytes).
No error found.
(For more information, see transcript file helloworld.log)

~/tralics$ cat helloworld.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE hello SYSTEM 'world.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<hello>
```

```
<p>Hello world!</p>
<p>Příliš žluťoučký kůň úpěl ďábelské ódy.
</p>
</hello>
```

Tralics nabízí několik nových maker a prostředí[7] pro definici struktury XML výstupu.

```
~/tralics$ vim helloworld.tex
            ...
~/tralics$ cat helloworld.tex
\documentclass{article}
\def\hello{\uppercase{h}ello}
\begin{document}
\hello{} \world!

\begin{xmlelement}{pokus0}
\begin{xmlelement}{pokus1}
Obsah prvního testovacího elementu.
\end{xmlelement}
\begin{xmlelement}{pokus2}
Obsah druhého testovacího elementu.
\end{xmlelement}
\AddAttToLast{attposlední}{hodnotaposlední}
\AddAttToCurrent{attsoučasný}{hodnotasoučasný}
\end{xmlelement}

\xbox{xboxelement}{Obsah \emph{xbox} elementu.}
\xmlelt{xmleltelement}{Obsah \emph{xmlelt} elementu.}

Příliš žluťoučký kůň úpěl ďábelské ódy.
\end{document}

~/tralics$ ./tralics -confdir=confdir/ -configfile=./hwconfig.tcf \
> -utf8 -utf8output helloworld.tex
            ...

~/tralics$ cat helloworld.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE hello SYSTEM 'world.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<hello>
<p>Hello world!</p>
<p><pokus0 attsoučasný='hodnotasoučasný'>
```

[7] Dokumentace je k dispozici na domovské stránce projektu (<http://www-sop.inria.fr/apics/tralics/>).

```
<pokus1>Obsah prvního testovacího elementu.
</pokus1>
<pokus2 attposlední='hodnotaposlední'>
Obsah druhého testovacího elementu.
</pokus2>
</pokus0></p>
<xboxelement>Obsah <hi rend='it'>xbox</hi> elementu.
</xboxelement><p><xmleltelement>Obsah <hi rend='it'>xmlelt</hi>
elementu.</xmleltelement></p>
<p>Příliš žluťoučký kůň úpěl ďábelské ódy.
</p>
</hello>
```

Matematické výrazy Tralics implicitně konvertuje do XML jazyka MathML.

```
~/tralics$ vim matematika.tex
           ...
~/tralics$ cat matematika.tex
\documentclass{article}
\begin{document}
Pythagorovu větu vyjadřuje tato rovnice: \[a^2 + b^2 = c^2\]
\end{document}

~/tralics$ ./tralics -confdir=confdir/ -utf8 -utf8output \
> -entnames=false matematika.tex
           ...

~/tralics$ cat matematika.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<std>
<p>Pythagorovu větu vyjadřuje tato rovnice:</p>
<formula type='display'>
  <math mode='display' xmlns='http://www.w3.org/1998/Math/MathML'>
    <mrow><msup><mi>a</mi> <mn>2</mn> </msup>
          <mo>+</mo>
          <msup><mi>b</mi> <mn>2</mn> </msup>
          <mo>=</mo>
          <msup><mi>c</mi> <mn>2</mn> </msup>
    </mrow>
  </math>
</formula>
<p noindent='true'/>
</std>
```

Argumentem `-nomathml` ale můžeme použití MathML potlačit. Matematické výrazy pak budou do dokumentu vkládány v podobě LaTeXového zdrojového textu výrazu.

```
~/tralics$ ./tralics -confdir=confdir/ -utf8 -utf8output \
> -entnames=false -nomathml matematika.tex
            ...

~/tralics$ cat matematika.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<std>
<p>Pythagorovu větu vyjadřuje tato rovnice:</p>
<texmath textype='display' type='display'>a^2 + b^2 = c^2</texmath>
<p noindent='true'/>
</std>
```

Pokud do dokumentu vložíme seznam literatury standardním LaTeXovým prostředím `thebibliography` bez podrobnějšího strukturního označkování obsahu, tak samozřejmě ani výsledek hlubší informaci o své struktuře neponese.

```
~/tralics$ vim matematika.tex
            ...
~/tralics$ cat matematika.tex
\documentclass{article}
\begin{document}
Pythagorovu větu vyjadřuje tato rovnice: \[a^2 + b^2 = c^2\]

Citovaná literatura: \cite{texbook,latex,chicago}.

\begin{thebibliography}{9}
\bibitem{texbook}
    Donald~E. Knuth.
    \textit{The \TeX book}.
    Addison-Wesley, 1984.
\bibitem{latex}
    Leslie Lamport.
    \textit{\LaTeX : A~Document Preparation System}.
    Addison-Wesley, 1986.
\bibitem{chicago}
    \textit{The Chicago Manual of Style},
    pages 400--401.
    University of Chicago Press, thirteenth edition, 1982.
\end{thebibliography}
\end{document}
```

```
~/tralics$ ./tralics -confdir=confdir/ -utf8 -utf8output \
> -entnames=false -nomathml matematika.tex
            ...

~/tralics$ cat matematika.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<std>
<p>Pythagorovu větu vyjadřuje tato rovnice:</p>
<texmath textype='display' type='display'>a^2 + b^2 = c^2</texmath>
<p>Citovaná literatura: <cit><ref target='bid0'/></cit>,
<cit><ref target='bid1'/></cit>, <cit><ref target='bid2'/></cit>.</p>
<Bibliography><p noindent='true'><bibitem id='bid0'/>Donald E. Knuth.
<hi rend='it'>The <TeX/>book</hi>.
Addison-Wesley, 1984.</p>
<p noindent='true'><bibitem id='bid1'/>Leslie Lamport.
<hi rend='it'><LaTeX/>: A Document Preparation System</hi>.
Addison-Wesley, 1986.</p>
<p noindent='true'><bibitem id='bid2'/>
<hi rend='it'>The Chicago Manual of Style</hi>,
pages 400-401.
University of Chicago Press, thirteenth edition, 1982.</p>
</Bibliography></std>
```

Tralics však dokáže pracovat také s bibliografickými databázemi programu BibTEX. Ty již informaci o struktuře záznamu obsahují a bude zachována i ve výstupním XML dokumentu.

Na rozdíl od překladu standardním LaTeXem se Tralics sám postará o zpracování bibliografické databáze. Není proto třeba volat externí programy nebo provádět několikanásobný překlad.

```
~/tralics$ vim matematika.tex
            ...
~/tralics$ cat matematika.tex
\documentclass{article}
\begin{document}
Pythagorovu větu vyjadřuje tato rovnice: \[a^2 + b^2 = c^2\]

Citovaná literatura: \cite{texbook,latex,chicago}.

\bibliographystyle{abbrv}
\bibliography{databaze-literatury}
\end{document}
```

```
~/tralics$ vim databaze-literatury.bib
            ...
~/tralics$ cat databaze-literatury.bib
@INBOOK{chicago,
   title = "The Chicago Manual of Style",
   publisher = "University of Chicago Press",
   edition = "Thirteenth",
   year = 1982,
   pages = "400--401",
   key = "Chicago"
}
@BOOK{texbook,
   author = "Donald E. Knuth",
   title= "The {{\TeX}book}",
   publisher = "Addison-Wesley",
   year = 1984
}
@BOOK{latex,
   author = "Leslie Lamport",
   title = "{\LaTeX}: {A} Document Preparation System",
   publisher = "Addison-Wesley",
   year = 1986
}

~/tralics$ ./tralics -confdir=confdir/ -utf8 -utf8output \
> -entnames=false -nomathml matematika.tex
~/tralics$ cat matematika.xml
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.6, date: 2010/02/18-->
<std>
<p>Pythagorovu větu vyjadřuje tato rovnice:</p>
<texmath textype='display' type='display'>a^2 + b^2 = c^2</texmath>
<p>Citovaná literatura: <cit><ref target='bid0'/></cit>,
<cit><ref target='bid1'/></cit>, <cit><ref target='bid2'/></cit>.</p>
<biblio>
<citation from='year' key='Chi82' id='bid2'
          userid='cite:chicago' type='inbook'>
<btitle>The Chicago Manual of Style</btitle>
<bedition>Thirteenth</bedition>
<bpublisher>University of Chicago Press</bpublisher>
<byear>1982</byear>
<bpages>400-401</bpages>
</citation>
```

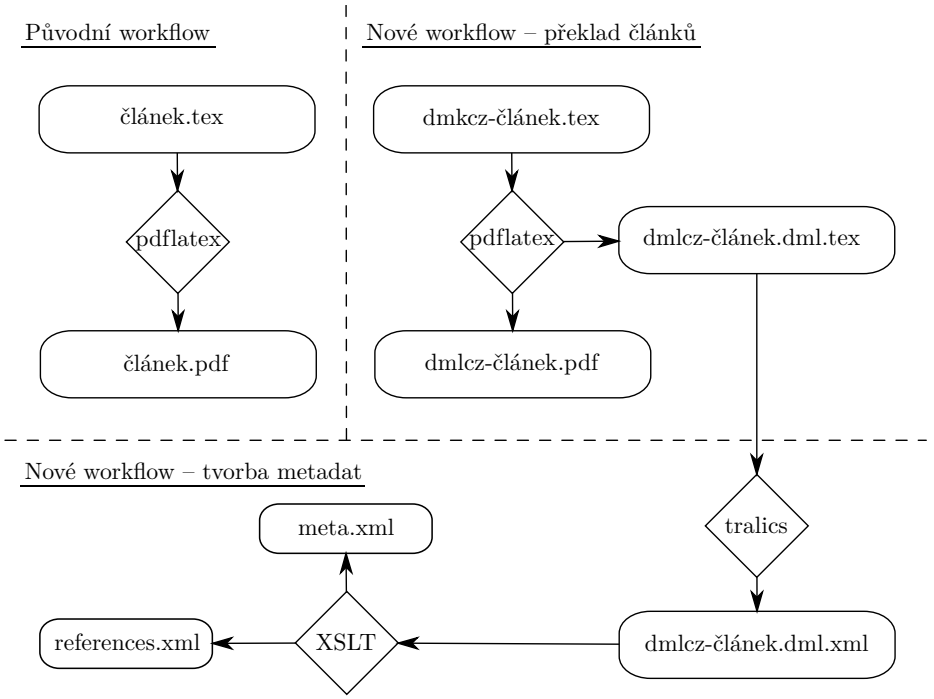
```
<citation from='year' key='Knu84' id='bid0'
          userid='cite:texbook' type='book'>
<bauteurs>
  <bpers prenom='D. E.' nom='Knuth' prenomcomplet='Donald E.'/>
</bauteurs>
<btitle>The <TeX/>book</btitle>
<bpublisher>Addison-Wesley</bpublisher>
<byear>1984</byear>
</citation>
<citation from='year' key='Lam86' id='bid1'
          userid='cite:latex' type='book'>
<bauteurs>
  <bpers prenom='L.' nom='Lamport' prenomcomplet='Leslie'/>
</bauteurs>
<btitle><LaTeX/>: A Document Preparation System</btitle>
<bpublisher>Addison-Wesley</bpublisher>
<byear>1986</byear>
</citation>
</biblio>
</std>
```

**Tralics v minimalistickém born-digital systém**

Jak bylo uvedeno výše, Tralics je také jádrem minimalistických born-digital systémů, jejichž cílem je co nejméně zasahovat do zvyklostí redakcí při jejich práci.

Jelikož je Tralics překladač LaTeXu do XML, bylo také vhodné vyhnout se závislosti na LaTeXu jako formátu používaného pro přípravu článků, neboť různé redakce mohou používat různé inkarnace TeXu. V případě minimalistických implementací proto byla modifikována původní redakční makra pracující s položkami, které tvoří součást DML-CZ metadat – název a autoři článků, abstrakty, seznamy klíčových slov, seznam literatury apod. Úprava byla provedena tak, že nemá vliv na výstupní podobu vysázeného textu, avšak argumenty maker jsou v průběhu překladu v původní podobě zapsány do pomocného souboru. (Navíc jsou obaleny makry Tralicsu identifikujícími význam té které položky.) Tamtéž zapisují parametry i některá nová makra, která pro potřeby generování DML-CZ metadat vznikla, i když nejsou přímo nezbytná pro sazbu časopisu (a nemají ani žádný viditelný efekt v sazbě) – jedná se např. o určení jazyka nebo typu článku apod.

Pokud redakce nepoužívají pro sazbu seznamů literatury BibTeX, největší rozšíření sady používaných maker se objevilo právě zde. Nová makra jsou určena pro vyznačení jednotlivých položek každého záznamu, neboť pro potřeby DML-CZ je třeba mít i seznamy citací co nejlépe strukturně označeny.




Obrázek 2: Schéma překladu článku bez použití Tralicsu a s použitím Tralicsu pro přípravu XML metadat

Při otevírání a uzavírání pomocného souboru jsou do něj doplněny LaTeXové úvodní a ukončující značky – tedy připojeny konkrétní textové řetězce, což je možné snadno provést i z jiných formátů než přímo z LaTeXu. Nejsme tedy vázaní jen na LaTeX jako původní formát článků, ale přesto může k vytvoření XML použít Tralics. Ve výsledku je obsahem pomocného souboru jednoduchý LaTeXový dokument využívající rozšířené sady maker Tralicsu, a obsahující strukturně označkované datové položky z původního článku. Tento dokument je tedy připraven pro zpracování Tralicsem.

S těmito úpravami může redakce své články zpracovávat identickým způsobem jako dříve. Při překladu jednotlivých článků však navíc vznikají pomocné soubory s informacemi o jednotlivých článcích, které jsou Tralicsem přeložitelné do XML. (Původní a novou situaci ilustruje obrázek 2.) Pro zjednodušení konfigurace nejsou generována přímo DML-CZ XML metadata, ale konečné XML soubory jsou získány aplikací XSLT.

Pokud jsou v položkách metadat použita makra definovaná přímo autorem článku, nebude jejich definice v pomocném souboru zpracovávaném Tralicsem obsažena. Tralics proto nebude vědět, jak s těmito makry naložit. Za tímto účelem bylo definováno speciální okolí, do kterého je možné v každém článku zapsat definice zde používaných uživatelských maker pro Tralics. Při překladu je obsah tohoto okolí uložen do souboru `\jobname.dml.ult`, který je Tralicsem při převodu pomocného souboru načítán.

## Závěr

V článku popsaný systém prokázal svou použitelnost v praxi – aktivně se používá v rámci projektu DML-CZ (odkud jsou data přebírána také do projektu Evropské digitální matematické knihovny EuDML). Rozsah užití je následující:

- Časopis *Archivum Mathematicum* vydávaný Masarykovou univerzitou (využívá komplexní systém od roku 2009),
- sborníky konference *DML* (*Towards a Digital Mathematics Library*) vydávané Masarykovou univerzitou (využívá přizpůsobený minimalistický systém od roku 2009),
- časopis *Communications in Mathematics* (do roku 2011 vydávaný pod jménem *Acta Mathematica et Informatica Universitatis Ostraviensis*) vydávaný Ostravskou univerzitou v Ostravě (využívá přizpůsobený minimalistický systém od roku 2010),
- časopis *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica* vydávaný Univerzitou Palackého v Olomouci (využívá přizpůsobený minimalistický systém od roku 2010),
- časopis *Kybernetika* vydávaný Ústavem teorie informace a automatizace Akademie věd České republiky (využívá přizpůsobený minimalistický systém od roku 2010).

Podobný systém byl použit také interně v rámci projektu DML-CZ při přípravě metadat publikací z retro-born-digital období, jmenovitě pro přípravu metadat

- osmi konferenčních sborníků *NAFSA* (*Nonlinear Analysis, Function Spaces and Applications*),
- časopisu *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica* od roku 2003,
- časopisu *Acta Mathematica et Informatica Universitatis Ostraviensis* od roku 2005,
- časopisu *Applications of Mathematics* (vydávaného Matematickým ústavem Akademie věd České republiky) od roku 1994,
- časopisu *Czechoslovak Mathematical Journal* (vydávaného Matematickým ústavem Akademie věd České republiky) od roku 1992,
- časopisu *Kybernetika* od roku 1998,
- časopisu *Mathematica Bohemica* (vydávaného Matematickým ústavem Akademie věd České republiky) od roku 1992,

- časopisu *Pokroky matematiky, fyziky a astronomie* (vydávaného Jednotou českých matematiků a fyziků) od roku 1993.

Doufáme, že popsané užití TEXových technologií nalezne uplatnění i u budoucích přispěvatelů projektu DML-CZ a popsané systémy tak budou nápomocny při jeho dalšímu rozvoji.

## Reference


[1] Miroslav Bartošek, Martin Lhoták, Jiří Rákosník, Petr Sojka, and Martin Šárfy. DML-CZ: The Objectives and the First Steps. In J. Borwein, M. Rocha E. and F. Rodrigues J. editors, *CMDE 2006: Communicating Mathematics in the Digital Era*, pages 69–79. A. K. Peters, MA, USA, 2008.

[2] Thierry Bouche. A pdfLATEX-based automated journal production system. *TUGboat*, 27(1):45–50, 2006.

[3] Centre de diffusion de revues académiques mathématiques. [online]. [cit. 2010-01-13]. URL: `<http://www.cedram.org/>`.

[4] Apics Team. Tralics: a LaTeX to XML translator. [online], October 2009. [cit. 2009-11-14]. URL: `<http://www-sop.inria.fr/apics/tralics/>`.

[5] Thierry Bouche. CEDRICS: When CEDRAM Meets Tralics. In Petr Sojka, editor, *Towards Digital Mathematics Library, Proceedings of the DML 2008 workshop*, pages 153–165.

[6] Eitan M. Gurari. TeX4ht: LaTeX and TeX for Hypertext. [online], June 2008. [cit. 2010-02-18]. URL: `<http://www.tug.org/applications/tex4ht/>`.

[7] CV Radhakrishnan. HCode : a web notebook extrapolating TeX4ht. [online], September 2009. [cit. 2010-02-18]. URL: `<http://www.tug.org/tex4ht/hcode.html>`.


## Summary: A Scientific Journal Processing System with the Capability of Exporting to a Digital Library


Production workflow of publishing scientific, especially mathematical journals is based on TEX and related technologies. Publisher usually prepare and make papers available electronically in a digital library, optimized for digital delivery and eventually for reading too.

Paper describes designed and implemented production workflow of several mathematical journals that archive their production in the Czech Digital Mathematics Library DML-CZ, which is subsequently available in the European Digital Mathematics Library EuDML.



*Masarykova univerzita, Fakulta informatiky, Botanická 68a, 602 00 Brno*
*<sojka@fi.muni.cz>, <xruzick7@fi.muni.cz>*