Zpravodaj Československého sdružení uživatelů TeXu

Jana Chlebíková

# Odborný dokument pre TeX a Web

*Zpravodaj Československého sdružení uživatelů TeXu*, Vol. 8 (1998), No. 3-4, 144–152

Persistent URL: http://dml.cz/dmlcz/149823

## Odkazy


[1] M. Goosens, S. Rahtz, F. Mittelbach, *The LaTeXGraphics Companion*. Addison Wesley 1997. ISBN 0-201-54199-8.

[2] K.M. Heal, M.L. Hansen, K.M. Rickard, *Maple V Learning Guide*. Springer-Verlag 1998. ISBN 0-387-98399-6.

[3] Petr Olšák, *Jak dostat obrázky z programu Mathematica do TeXu*. Zpravodaj Československého sdružení uživatelů TeXu, **3** (1), 34–40 (1993).



*Roman Plch*
plch@math.muni.cz


# Odborný dokument pre TeX[1] a Web


Janka Chlebíková


Nasledujúci príspevok je počítačovou sondou do odborného (predovšetkým matematického) dokumentu bez ohľadu na hĺbku odborných výsledkov v ňom obsiahnutých...

Nástup osobných počítačov a počítačových sietí priniesol výrazné zmeny v súvislosti s odborným dokumentom. Niekoľko storočí existujúcu papierovú formu dokumentu (resp. mikrofiše) dopĺňa nová elektronická forma dokumentu s rôznymi formátmi pre uchovávanie (Postscript, PDF, HTML a ďalšie) na rôzne typy médií.

Elektronická forma pridáva dokumentom nové rozmery (napr. „živé referencie“, multimediálne prvky, či vyhľadávanie v dokumente), čo v spojení s novými počítačovými technológiami znamená predovšetkým zmenu v *sprístupnení* a *multifunkčnom* využití odborného dokumentu [2].

Zmena zasiahla aj priamo proces vytvárania tlačenej podoby dokumentu. Počítače, tlačiarne, či osvitové jednotky s kvalitným softvérom takmer úplne vytlačili klasické sádzacie stroje.

## 1. TeX a odborné dokumenty

Počítačová sadzba odborných dokumentov sa takmer od jej vzniku nesie jednoznačne v znamení TeXu. Napriek tomu, že sa neustále vyvíjajú čoraz kvalitnejšie

[1] Pre naše účely TeX ∼ LaTeX ∼ $\mathcal{AMS}$-TeX.

DTP systémy, TEX ako typografický systém pre náročné odborné dokumenty ostáva stále na stupni víťazov.

Navyše TEXový jazyk je najrozšírenejším jazykom v elektronickej komunikácii medzi matematikmi (e-mail, news-groupy, mailing listy), ktorú so sebou priniesol Internet. TEX pre svoju platformovú nezávislosť má svoj leví podiel aj na jednoduchom *sprístupňovaní* odborných dokumentov (napr. výmena elektronických dokumentov).

## 2. Odborný dokument a súčasný Web

Internet však priniesol svetu i populárny Web spolu s HTML jazykom. V tomto jazyku je viditeľná snaha o vytvorenie dostatočného množstva značiek (tzv. tagov) na vyjadrenie štruktúry univerzálneho typu dokumentu (univerzálny dokument = odborný matematický článok, dopis, slide, ...). Z tagov potrebných pre popis štruktúry samotného matematického výrazu je podporovaný len index a exponent. Akékoľvek zložitejšie matematické formulky predsa typický užívateľ nepoužíva a nieto aby ich ešte zverejňoval na Webe:-)

Stručne povedané, jazyk HTML neposkytuje dostatočné množstvo tagov pre popis štruktúry matematických výrazov. Navyše HTML má fixnú sadu tagov a tak si ich užívateľ nemôže sám pridávať podľa potreby. Na druhej strane, Web je najpopulárnejší prostriedok na *sprístupňovanie* dokumentov. Z tohto dôvodu sa venuje pozornosť rôznym prístupom zobrazovania odborných dokumentov na Webe. Ich spoločnou črtou je:

- dôraz len na primárny cieľ zobrazenia matematického výrazu na Webe
- v podstate všetky predpokladajú vstupné kódovanie v TEXovom jazyku

### 2.1. Riešenia založené na báze HTML

Uveďme niekoľko najbežnejších prístupov:

1. Najrozšírenejším prístupom je preklad matematických výrazov, ktoré sa nedajú v HTML popísať, do obrázkov (najčastejší formát GIF) a tie zobraziť ako súčasť dokumentu. Takýto prístup ale prináša rad nevýhod:

- je potrebný ľudský faktor na zachytenie kontextu obrázku, nie je možné vyhľadávanie v takýchto obrázkoch, resp. akékoľvek využitie ich obsahu,
- pri zmene veľkosti fontu sa veľkosť obrázku nemení,
- problém s vhodným riadkovaním, odsadzovanie za sebou idúcich riadkov,
- kvalita tlačenia je závislá od kvality obrázkov a nezodpovedá kvalite vytlačeného textu okolo výrazu,
- pomalé načítavanie stránok s množstvom obrázkov.

Príklady implementácií: LaTeX2HTML [`http://www-dsed.llnl.gov/files/programs/unix/latex2html`], TeX4ht [`http://www.cis.ohio-state.edu/~gurari/TeX4ht/mn.html`].

2. Substitúcia TEXových elementov do symbolov z fontov dostupných bežne na každom počítači (napr. Symbol font). Každý výraz je prekladaný do štruktúry, ktorá môže byť popísaná priamo v HTML jazyku s použitím „bežných“ symbolov. Z dôvodu rýchleho zobrazenia (ale zlej kvality) je takýto prístup vhodný len na získanie prvej informácie o obsahu webovej stránky.

Príklady implementácií: TeX2HTML [`http://www.tex2html.com`].

3. Zobrazovanie matematických výrazov je zabezpečované cez rôzne applety, či plug-iny. Vstupné kódovanie v tomto prípade nie je viazané len na TEXový jazyk, ale rôzne webové nadstavby sú schopné zobraziť aj iné vstupy. Nevýhodou tohto prístupu je závislosť na konkrétnom browseri, či platforme, ktorý je schopný zobraziť vstupné kódovanie. Často je nutnosťou inštalácia podporných programov a potrebný dlhý čas na zobrazenie dokumentu.

Uveďme niektoré implementácie tohto prístupu:

- IBM techexplorer [`http://www.ics.raleigh.ibm.com`]. Veľmi zaujímavé plug-iny pre IE (Internet Explorer) a Netscape Navigator na najrozšírenejšie platformy. Vedia priamo interpretovať podmnožinu TEXových, (LaTEXových) príkazov a cez plug-in zobraziť takéto dokumenty na Webe. V jednom zo svojich projektov realizovali aj prepojenie elektronických dokumentov s computer algebra systémom AXIOM.
- WebEQ Equation Rendering [`http://www.geom.umn.edu/~rminer/jmath`]. Java applet, ktorý zobrazuje matematické výrazy vložené cez embedded elementy (LATEXový popis) do HTML kódu.

### 2.2. Riešenia bez HTML – Web je iba sprostredkovateľ

V súčasnosti najrozšírenejším spôsobom pre sprístupňovanie odborných dokumentov je použitie niektorého zo štandardných elektronických formátov a Web len ako informačný sprostredkovateľ. Najčastejšie sú používané formáty PS (Postscript), PDF (Portable Document Format), DVI (Device Independent). Takouto formou dnes publikujú na Webe desiatky odborných časopisov [`http://www.emis.de/journals`]. Výhodou tohto prístupu je kontrola autora dokumentu nad konečným vzhľadom dokumentu, čo pri použití HTML jazyka zďaleka nie je možné. Nevýhodou je ale veľkosť prenášaných súborov, nutnosť inštalácie prehliadačov pre každý formát a v niektorých prípadoch problémy s nekompatibilitou formátov.

## 3. Odborný dokument a budúci Web

Konzorcium w3c [`http://www.w3.org`] zodpovedné za vývoj štandardov na Webe sa po dlhých úvahách rozhodlo uvoľniť ohraničenosť HTML jazyka. Prijatie XML (Extensible Markup Language) vo februári 1998 znamená pre uží-

vateľov predovšetkým možnosť pridávania vlastných tagov, resp. tagov podľa zvoleného DTD (Document Type Definition). XML je totiž „mladším bratom“ SGML a nielen jedným z mnohých DTD, ako tomu bolo v prípade HTML. Jedno výstižné prirovnanie z XML FAQ „XML je skôr SGML-- ako HTML++“.

Prijatím XML sa otvoril nový priestor pre odborné dokumenty na Webe. MathML (Mathematical Markup Language) ako nové DTD v XML poskytuje nový štandard pre kódovanie (značkovanie) matematických výrazov na Webe [`http://www.w3.org/MathML`].

Prirodzená otázka je: na základe akej prezentácie bude browser zobrazovať nové elementy[2], nakoľko XML je jazyk len na popis štruktúry dokumentu. V HTML jazyku je situácia jednoduchšia, lebo fixná množina tagov je dopredu známa. Browser tak môže mať v sebe predpísanú prezentáciu jednotlivých elementov (veľkosť a typ fontov, veľkosť odsadenia, ...). Tento fakt bol ale príliš obmedzujúci a preto bola v HTML jazyku pridaná možnosť kontroly niektorých prezentačných vlastností elementov pomocou CSS (Cascading Style Sheets). Podstatne silnejšie možnosti prezentačnej kontroly poskytuje XSL (Extensible Style Language), ktorý je navrhnutý ako nový štandard pre popis prezentácie XML dokumentov.

Je len otázkou času, kedy dva najrozšírenejšie browsery Netscape a IE začnú „rozumne“ podporovať XML. (Nejaká malá podpora už existuje aj dnes, posledná verzia IE nezobrazuje XML tagy.) Zobrazovanie matematických výrazov na Webe by mohlo byť potom rovnako prirodzené a jednoduché, ako je to v súčasnosti s textom bez nich. Že nerozprávame o príliš vzdialených métach je možné presvedčiť sa na browseri Amaya (experimentálny browser w3c), ktorý podporuje zobrazovanie a jednoduché štrukturované (WYSIWYG) editovanie časti MathML tagov [`http://www.w3.org/Amaya/`].

Vráťme sa však teraz k MathML a pozrime sa bližšie na jeho ciele a možnosti, ktoré poskytuje.

### 3.1. Ciele MathML

Základným cieľom MathML je, aby matematické výrazy mohli byť podávané, obdržané a spracované na Webe práve tak, ako je to možné pri použití HTML s jednoduchým textom.

Podstatný rozdiel od doterajších prístupov je, že dôraz je kladený nielen na *sprístupnenie* dokumentu, t.j. jeho „pekné“ zobrazenie cez Web, ale i na *multifunkčnom* využití obsahu matematických výrazov (automatické spracovávanie, vyhľadávanie, indexovanie, ..., resp. prepojenie na computer algebra systémy).

Voľba vhodného kódovania dokumentu je dôležitá pre životnosť dokumentov. Vzhľadom k tomuto aspektu MathML je navrhnutý tak, aby umožňoval kódovať

[2]Element je štrukturálna časť dokumentu uzavretá medzi dvoma párovými tagmi, napr. `<LI>...</LI>`.

nielen štruktúru matematického výrazu (syntax), ale i jeho sémantickú časť. Takto je možné realizovať prepojenie s ďalšími aplikáciami.

Dokument označkovaný MathML tagmi je „ľudsky čitateľný“, čo umožňuje jeho jednoduché generovanie z a do iných systémov (napríklad komunikácia s TEXom), automatické spracovanie softvérom, či v krajnom prípade priame editovanie v jednoduchých textových editoroch. (V každom prípade však nie je určený pre „ručné“ editovanie!)

MathML je navrhnutý tak, aby umožňoval pridanie informácie pre špeciálne prehliadače (napríklad zvukový výstup výrazu pre handicapovaných) a rôzne aplikácie. Pokrýva všetky existujúce matematické materiály vhodné pre vedecké a študijné účely.

### 3.2. Čo vplývalo na MathML

V nemalej miere to bol TEX, ktorý je však prezentačným jazykom a preto nedostatočný pre multifunkčné ciele MathML dokumentov. Z dôvodu premenlivej veľkosti okna a fontu, je potreba i častého zalamovania matematického výrazu. MathML jazyk musí poskytovať dostatočné prostriedky na špecifikáciu miest riadkového zlomu.

Na MathML vplývalo aj ISO 12083 Maths DTD, ktoré však zachytáva tiež viac prezentačné vlastnosti matematických výrazov ako sémantické.

Na sémantickú časť MathML výraznou mierou vplýval Open Math, zastrešujúci computer algebra systémy ako Mathematica, či Maple, ako i samostatné zmienené systémy.

## 4. Všeobecné princípy MathML

MathML pozostáva z dvoch samostatných skupín tagov: *prezentačných* a *obsahových*. Skupiny sú navzájom nezávislé (t.j. výraz sa popisuje buď v prezentačných, alebo obsahových tagoch), ale za istých pravidiel sa tagy oboch skupín môžu miešať.

### 4.1. Prezentačné tagy

Cieľom *prezentačných* tagov je popísať štruktúru matematických výrazov tak, aby bolo možné dosiahnuť vysoko kvalitný výstup na výstupných zariadeniach: obrazovke či tlačiarni. Sú v podstate analógiou TEXu s tým rozdielom, že užívateľ je nútený presne implicitne špecifikovať čo je operátor, relácia, premenná, či číslo. Odpovedajú 2-dimenzionálnemu označeniu výrazov typu zlomok, exponent, index, či matica. Atribúty môžu nepriamo ovplyvňovať prezentáciu, napríklad či zátvorky majú vertikálne prekrývať obsahujúci výraz, predpisovať minimálnu veľkosť fontu pri indexoch a podobne.

Príklad použitia prezentačných tagov pre výraz: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

```
<mrow>
  <mi>x</mi>
  <mo>=</mo>
  <mfrac>
    <mrow>
      <mrow> <mo>-</mo> <mi>b</mi> </mrow>
      <mo>&PlusMinus;</mo>
      <msqrt>
        <mrow>
          <msup> <mi>b</mi> <mn>2</mn> </msup>
          <mo>-</mo>
          <mrow>
            <mn>4</mn>
            <mo>&InvisibleTimes;</mo>
            <mi>a</mi>
            <mo>&InvisibleTimes;</mo>
            <mi>c</mi>
          </mrow>
        </mrow>
      </msqrt>
    </mrow>
    <mrow> <mn>2</mn> <mo>&InvisibleTimes;</mo>
           <mi>a</mi>
    </mrow>
  </mfrac>
</mrow>
```

### 4.2. Obsahové tagy

*Obsahové* tagy vyjadrujú sémantiku výrazu pomocou prostriedkov daných v príslušnej matematickej oblasti (teória množín, lineárna algebra, ...). Uveďme jednoduchý príklad pre porovanie s prezentačnými tagmi. Pokiaľ pre prezentáciu výrazu $x^2$, je postačujúci zápis:

```
<mrow> <msup> <mi> x</mi> <mo> 2</mo></msup></mrow>
```

tak pre jeho sémantiku je nutné vyjadrenie, že horný index znamená umocnenie na príslušný exponent:

```
<expr> <mi> x </mi> <power/> <mn> 2</mn> </expr>
```

Obsahové tagy zahŕňajú širokú škálu prázdnych elementov pre operátory, relácie a pomenované funkcie, ako napríklad `<plus/>`, `<leq/>`, či `<tan/>`.

Umožňujú tiež vyznačiť konštruktor univerzálnej funkcie a jej argumenty, ako i analogické sémantické informácie potrebné na prepojenie napríklad na computer algebra systémy.

```
<apply> <minus/> <ci>a</ci> <ci>b</ci> </apply>
```

Pre porovnanie uveďme opäť zápis $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ v obsahových tagoch:

```
<reln>
  <eq/>
  <ci>x</ci>
  <apply>
    <over/>
    <apply>
      <fn occurrence="infix"><mo>&PlusMinus;</mo></fn>
      <apply> <minus/> <ci>b</ci> </apply>
      <apply>
        <root/>
        <apply>
          <minus/>
          <apply>
            <power/> <ci>b</ci> <cn>2</cn>
          </apply>
          <apply>
            <times/> <cn>4</cn> <ci>a</ci> <ci>c</ci>
          <apply>
        </apply>
        <cn>2</cn>
      </apply>
    </apply>
    <apply>
      <times/> <cn>2</cn> <ci>a</ci>
    </apply>
  </apply>
</reln>
```

### 4.3. Miešanie obsahových a prezentačných tagov

Rozhodnutie, či použiť obsahové alebo prezentačné tagy, závisí od účelu využitia dokumentu. Ak sú to študijné materiály s perespektívou ich ďalšieho využitia, je lepšie použiť obsahové tagy. Ak nie sú vhodné sémantické prostriedky (napríklad pre $\pm$ v príklade), či pre nové smery v matematike, môžeme použiť len prezentačné tagy

V spomínanom kontexte je vhodné spomenúť ešte jeden veľmi perespektívny element `semantics`. Používa sa na vyjadrenie časti štruktúry matematického výrazu nielen v MathML jazyku, ale i rôznych iných druhoch kódovania (vyjadrené atribútom). Tieto kódovania nemusia mať nič spoločné s XML, ale efektívne môžu byť využité (TEX, jazyk C, Maple) pri iných aplikáciach. Napríklad časť výrazu môže byť vyjadrená v prezentačných a zároveň i v obsahových tagoch, alebo v kódovaní v jazyku Maple, či v TEXu. Takýmto spôsobom sa rozširujú sémantické možnosti obsahových MathML tagov.

Jednoduchý príklad:

```
<semantics>
  <apply><plus/>
    <apply><sin/> <ci> x </ci> </apply>
    <cn> 5 </cn>
  </apply>
  <annotation encoding="TeX">
    \sin x + 5
  </annotation>
</semantics>
```

### 4.4. Ako je to s matematickými symbolmi?

XML podporuje UNICODE, čím je vytvorený priestor pre podporu veľkého množstva štandardizovaných matematických symbolov. MathML entity obsahujú funkčné pomenovania najpoužívanejších matematických symbolov z UNICODE (napríklad entita `&PlusMinus;` z príkladu).

V rámci projektu Stick je snaha o zjednotenie matematických symbolov a fontov používaných v TEXu, podľa ISO noriem, v UNICODE, computer algebra systémoch, či jednotlivými odbornými vydavateľstvami (napr. Elsevier).

## 5. Podpora MathML v súčasnosti

Dnes je už MathML podporované viacerými softvérovými produktami. Spomeňme aspoň najdôležitejšie:

- Computer algebra systémy Maple [`http://www.maplesoft.com/`] či Mathematica [`http://www.wolfram.com/news/mathml/`] sú schopné exportovať, importovať a vyhodnotiť MathML výrazy (v obsahových tagoch).
- IBM techexplorer cez plug-in zobraziť v IE alebo Netscape časť MathML tagov.
- WebEQ cez Java applet vie zobraziť v HTML dokumente MathML tagy.

Samostatnú časť je možné venovať vytváraniu MathML dokumentov.

### 5.1. Ako vytvárať odborné dokumenty pre Web?

TEXové dokumenty je bežné vytvárať „ručne“. TEX má totiž vysoký stupeň inteligencie a rozozná napríklad operátor od relácie. Tým pádom potrebujeme značku (riadiace slovo) vložiť len pri zmene prezentácie, resp. popisovaní dvojrozmernej štruktúry výrazu, teda optimalizovaný počet krát.

V MathML je nutné značkovať všetky štrukturálne elementy už na najnižšej úrovni, napríklad čísla a premenné. Ako vidieť aj z príkladov, v prípade MathML dokumentov by „ručné“ vytváranie bolo krokom späť. Preto je veľká pozornosť venovaná aj problematike vytvárania takto kódovaných dokumentov.

Pohodlným spôsobom možno MathML dokumenty vytvárať v štrukturovaných editoroch [1]. Ich výhodou je, že vytváraný dokument vždy zodpovedá zvolenému DTD. Umožňujú prípadne jednoduchý export do TEXu ako typografickému systému pre účely tlače. Pre experimentovanie so štrukturovaným editovaním možno doporučiť Amayu, či Euromath systém [`http://www.dcs.fmph.uniba.sk/~emt`].

V súčasnosti je rozbehnutých niekoľko nezávislých projektov pre automatické generovanie MathML dokumentov z TEXových a naopak. Nakoľko TEX je prezentačný jazyk, jedná sa o preklad len do prezentačných tagov. Tento prístup je ale veľmi dôležitý, pretože dnes naprostá väčšina matematických dokumentov ak je v elektronickej podobe, tak je v TEXu. Navyše TEX je prirodzeným jazykom matematikov pre ústnu i písomnú podobu a dá sa očakávať, že ho v dohľadnej dobe nič iné nenahradí. Spomínaný prístup je teda výhodný pre priamočiare sprístupnenie TEXových dokumentov na Webe.

## 6. Literatúra

[1] Janka Chlebíková, *Štrukturované editovanie.* Zpravodaj Československého sdružení uživatelů TEXu, **7** (4), 185–190 (1997).

[2] Philip Taylor, *Computer typesetting or electronic publishing? New trends in scientific publication.* Zpravodaj Československého sdružení uživatelů TEXu, **5** (1–4), 61–89 (1995).

*Janka Chlebíková*
*Katedra vyučovania informatiky*
*MFF UK, Bratislava*
`chlebikj@dcs.fmph.uniba.sk`