# XML jako wejście Omegi

Paweł Grams
ENST-Bretagne BP 832
29-285 Brest CEDEX, France
pawel.grams@enst-bretagne.fr

Pracę zgłosił: Andrzej Borzyszkowski

**Streszczenie**

Obecne publikacje opracowane za pomocą TEXa lub LATEXa rzadko są przekształcane na postać hipertekstową. Ponadto komplikacja języka TEX, a w szczególności skomplikowana i zależna od kontekstu składnia, sprawiają iż trudno jest stworzyć oprogramowanie do automatycznej obróbki danych w tym formacie. Co prawda możliwa jest zmiana formatu LATEX na HTML i odwrotnie, lecz wymaga to odpowiedniego sformatowania pliku LATEX, co jednocześnie wiąże się z ograniczeniem swobody autora podczas tworzenia struktury dokumentu.

## Wprowadzenie

TEX jest językiem programowania dającym doskonałe możliwości formatowania tekstu. Jednakże stopień jego komplikacji czyni go niedostępnym dla początkującego użytkownika. Głównie z tego powodu powstał LATEX, udostępniający użytkownikowi zestaw prostych makr, pozwalających w pełni wykorzystać możliwości TEXa przy minimum wiedzy o składaniu tekstu. System Omega również daje takie możliwości, jednocześnie pozwalając na obsługę wszystkich znaków Unicode. Kierunkiem dalszej ewolucji jest bardziej zunifikowana wersja LATEX3.

Tematem mojego stażu, prowadzonego na École Nationale Supérieure des Télécommunications de Bretagne pod nadzorem prof. Yannisa Haralambousa, jest zaadaptowanie Omegi do środowiska XML poprzez nadanie jej możliwości czytania plików w tym formacie. Docelowo Omega będzie miała również zdolność zapisywania w plikach wyjściowych nie tylko jak dotychczas danych DVI, lecz również dodatkowych, redundantnych danych w formatach takich jak SVG, MathXML lub inne. Pozwoli to na efektywne przetwarzanie danych wejściowych za pomocą istniejących narzędzi ze świata XML.

## Realizacja projektu

Główną przeszkodą utrudniającą stworzenie spójnego zestawu znaczników XML na podstawie komend LATEXa jest mieszanie w tym ostatnim, zawartości merytorycznej dokumentu z danymi służącymi do formatowania tekstu. Jako przykład niech posłuży komenda:

```
\begin{tabular}{c|c|c}
```

Wyraźnie widać, że w nawiasach klamrowych mamy tylko i wyłącznie informacje potrzebne do poprawnego ułożenia tekstu na stronie. Natomiast w komendzie `\emph{tekst}` tekst w nawiasach zawiera treści dokumentu. Najtrudniejsze do obsługi są komendy gdzie tekst dokumentu miesza się z parametrami niezbędnymi do składu:

```
\begin{tabular}{r@{równe}l}
```

Taka konstrukcja komend LATEX bardzo utrudnia automatyczną konwersję danych wejściowych na inny format, zautomatyzowane sprawdzanie poprawności plików, oraz konstrukcję narzędzi do wstępnej obróbki danych.

Przy tworzeniu XLATEX starałem się stworzyć znaczniki XML podobne w konstrukcji do odpowiadających im komend LATEXa. Miało to na celu ułatwienie korzystania z nowego narzędzia osobom które już używały LATEXa. Pozostawiłem również wiele środowisk w ich pierwotnej formie. Można się do nich odwołać za pomocą znaczników typu: `<?math tekst ?>`. Wewnątrz takich znaczników używać można tylko komend (LA)TEXa dozwolonych dla danego środowiska. Miało to na celu zmniejszenie liczby znaczników XML do niezbędnego minimum, które pozwoli na wprowadzenie LATEXa do świata XML, a jednocześnie nie będzie wymagało od użytkowników czasochłonnego zgłębiania listy wszystkich dostępnych znaczników.

## Strona techniczna

Ze strony technicznej, pierwsza część przedsięwzięcia została zaprojektowana tak, aby pozwolić na łatwą adaptację do późniejszych zmian wprowadzonych w strukturze XLATEX.

Mianowicie jedyna zmiana w kodzie Omegi polega na wbudowaniu możliwości korzystania z zewnętrznego filtra wstępnie przetwarzającego dane wejściowe. Filtr ten, pobierał będzie dane z pliku, konwertował je na format LaTeX i na bieżąco przekazywał do Omegi. Ze strony użytkownika wymagane będzie tylko podanie w linii komend Omegi nazwy pliku zawierającego filtr niezbędny do poprawnego odczytania pliku wejściowego. Użytkownik korzystający z formatu XLaTeX, nie będzie miał więc w ogóle do czynienia z plikami w formacie LaTeX.

Zastosowanie filtru pozwoli zaawansowanym użytkownikom zainteresowanym zmianami w formacie XLaTeX lub wprowadzeniem nowego formatu, na łatwą adaptację Omegi do swoich potrzeb.

## Struktura XLaTeX

Podczas tworzenia struktury znaczników XLaTeX starałem się rozdzielić treść dokumentu od parametrów komend LaTeX, definiując te ostatnie jako wartości argumentów elementów XML. Jedynym wyjątkiem, gdzie nie udało się uniknąć pomieszania tych dwóch typów danych jest środowisko `tabular`. Jak wiadomo, w definicji tabeli, za pomocą komendy `@{}` można wstawić dowolny tekst, który następnie będzie systematycznie umieszczany w tym miejscu, w każdym wierszu. Aby uniknąć komplikacji struktury XLaTeX zaproponowałem dwie wersje znacznika `<tabular>`. W pierwszej, format tabeli podawany jest bezpośrednio jako wartość argumentu, zaś w drugiej, jest on objęty znacznikami `<format>`.

Jeśli chodzi o wartości argumentów, powinny one dokładnie odpowiadać wartościom jakie przyjmują parametry odpowiednich komend języka LaTeX. Nie mogą one zawierać znaczników XLaTeX. Trzeba również pamiętać o tym, iż argumenty obowiązkowe dla komend LaTeX w XLaTeX również pozostają argumentami, których obecność jest obowiązkowa.

Jedynym dodanym sztucznie argumentem jest `star`. Został on wprowadzony w celu zmniejszenia ilości znaczników XLaTeX. Jeśli jego wartością jest `yes`, otrzymujemy środowisko „z gwiazdką”. Jeśli argument jest nieobecny, lub różny od `yes`, wówczas otrzymamy środowisko bez gwiazdki.

Należy zauważyć iż podczas zmiany formatu plików z TeX na XLaTeX, zmienił się również zestaw znaków, które należy chronić. Obecnie należy chronić znaki `<` `>` `&` w całym pliku, oraz `"` `'` wewnątrz wartości argumentów.

## Środowiska

W XLaTeX lista znaczników XML obejmuje wiele podstawowych komend niezbędnych do napisania poprawnego dokumentu. Pozostałe komendy można uzyskać korzystając bezpośrednio z możliwości wpisywania kodu TeX. Taką możliwość daje instrukcja zmiany trybu: `<?tex >`. Postanowiliśmy jednak wprowadzić dodatkowo również inne tryby, odpowiadające często używanym środowiskom.

W większości przypadków przejście wymaga od filtru jedynie użycia istniejącego środowiska. Jednakże w przypadku środowisk `verbatim` i `verbatim*` zdecydowaliśmy się na stworzenie nowego środowiska, które będzie mogło być używane bez ograniczeń wewnątrz innych środowisk. W tym celu, filtr dodaje[1] automatycznie na początek pliku definicję nowego środowiska: `\newenvironment{my}{\begin{alltt}}{\end{alltt}}`. Następnie wszystkie znaki `\` obecne wewnątrz, są zamieniane na `\textbackslash`, oraz wprowadzana jest ochrona nawiasów klamrowych. W efekcie otrzymujemy tekst identyczny z tym produkowanym przez środowisko `verbatim`, lecz możliwy do uzyskania w każdych warunkach.

Środowisko `verbatim*` jest możliwe do uzyskania poprzez użycie trybu:

```
<?verbatim star="yes" >
```

który daje identyczny efekt. Zasada działania jest taka sama jak dla `verbatim`, z tym iż dodatkowo spacje zamieniane są na `\vs` zdefiniowane jako:
`\def\vs{\leavevmode\hbox{\tt\char`\ }}` i odpowiadające dokładnie znakowi używanemu do zaznaczania spacji w środowisku `verbatim*` Definicja ta jest dodawana[1] automatycznie przez filtr na początku pliku.

## Zakończenie

Celem tej publikacji, jak i wystąpienia na konferencji BachoTeX 2003, jest uzyskanie odzewu ze strony środowiska TeX-owego, na temat poprawności ustalonego przez nas zestawu znaczników XLaTeX. Chcielibyśmy w niedalekiej przyszłości ogłosić ostateczny zestaw znaczników XLaTeX, w postaci użytecznej i niewymagającej poważnych zmian. Kolejnym krokiem będzie nadanie Omedze możliwości zapisywania dodatkowych danych w plikach wyjściowych, również w postaci znaczników XML (ewolucja formatu DVI do DVX). Być może uda się uzyskać zapis dodatkowej porcji danych w formatach SVG, MathXML lub innych podobnych. Taki kierunek ewolucji Omegi pozwoli na zwiększenie możliwości obróbki tekstu, zarówno w formie jeszcze nie przetworzonej, jak i już gotowej do druku.

[1] dodaje virtualnie, gdyż Omega nigdzie nie zapisuje pliku tworzonego przez wyjście filtru.

### Lista markerów XML zdefiniowana do dnia 02–04–2003

| | |
|---|---|
| `documentclass, usepackage` | `<documentclass options="options" name="name"/>`<br>$\Longrightarrow$ `\documentclass[options]{class}` |
| `include, includeonly, input` | `<include file="file"/>` $\Longrightarrow$ `\include{file}` |
| `pagestyle, thispagestyle` | `<pagestyle name="name"/>` $\Longrightarrow$ `\pagestyle{name}` |
| `table` | `<table pos="pos" star="yes"> text </table>`<br>$\Longrightarrow$ `\begin{table*}[pos] text \end{table*}` |
| `letter` | `<letter><recipient> text </recipient>`<br>`maintext </letter>`<br>$\Longrightarrow$ `\begin{letter}{text} maintext \end{letter}` |
| `tabular`<br>Pomiędzy znacznikami `format` można stosować inne znaczniki XLaTeX | `<tabular format="format"> text </tabular>`<br>lub:<br>`<tabular><format> format </format>`<br>`text </tabular>`<br>$\Longrightarrow$ `\begin{tabular}{format} text \end{tabular}` |
| `chapter, paragraph, part, subparagraph, section, subsection, subsubsection` | `<section star="yes"> text </section>` $\Longrightarrow$ `\section*{text}`<br>lub:<br>`<section><header> text1 </header><toc>`<br>`text2 </toc> text </section>`<br>$\Longrightarrow$ `\section[text1]{text}\tocsection{text2}` |
| `linebreak, nolinebreak, nopagebreak, pagebreak` | `<linebreak weight="weight"/>`<br>$\Longrightarrow$ `\linebreak[weight]` |
| `address, date, signature` | `<address> text <address/>` $\Longrightarrow$ `\address{text}` |
| `glossary, index`<br>Argument `output` powinien zawierać komendy LaTeX niezbędne do poprawnego wyświetlenia tekstu | `<index entry="entry/>"` $\Longrightarrow$ `\index{entry}`<br>lub:<br>`<index output="output"> text`<br>`<index output="output1"> text1 </index></index>`<br>$\Longrightarrow$ `\index{text@output!text1@output1}` |
| `multicolumn` | `<multicolumn cols="colsnumber" pos="pos">`<br>`text </multicolumn>`<br>$\Longrightarrow$ `\multicolumn{colsnumber}{pos}{text}` |
| `cline` | `<cline range="range"/>` $\Longrightarrow$ `\cline{range}` |
| `item` | `<item mark="mark"> text </item>`<br>$\Longrightarrow$ `\item[mark] text` |
| `hyphenation` | `<hyphenation words="words"/>`<br>$\Longrightarrow$ `\hyphenation{words}` |
| `group` | `<group> text </group>` $\Longrightarrow$ `{text}` |
| `abstract, center, description, document, enumerate, flushleft, flushright, itemize, quotation, quote, tabbing, titlepage, trivlist, verse` | `<document> text </document>`<br>$\Longrightarrow$ `\ begin{document} text \end{document}` |
| `caption` | `<caption star="yes"> text </caption>`<br>$\Longrightarrow$ `\caption*{text}`<br>lub:<br>`<caption><header> text1 </header> text </caption>`<br>$\Longrightarrow$ `\caption[text1]{text}` |
| `author, cc, closing, date, displaylines, encl, fbox, hbox, mbox, opening, ps, thanks, title, vbox` | `<mbox> text </mbox>` $\Longrightarrow$ `\mbox{text}` |
| `and, appendix, bigskip, cleardoublepage, clearpage, dotfill, fussy, hline, hrulefill, indent, kill, ldots, listoffigures, listoftables, makeglossary, makeindex, maketitle, medskip, newline, newpage, noindent, sloppy, smallskip, tableofcontents, vline` | `<newpage/>` $\Longrightarrow$ `\newpage` |

| | |
|---|---|
| bf, emph, footnotesize, huge, Huge, it, large, Large, LARGE, md, normal, normalsize, rm, sc, scriptsize, sf, sl, small, tiny, tt, up | <rm> text </rm> $\Longrightarrow$ \begin{rm} text \end{rm} |
| footnote | <footnote id="id"> text </footnote><br>$\Longrightarrow$ \footnote[id]{text} |
| footnotemark | <footnotemark id="id"/> $\Longrightarrow$ \footnotemark[id] |
| footnotetext | <footnotetext id="id"> text </footnotetext><br>$\Longrightarrow$ \footnotetext[id]{text} |
| figure | <figure pos="pos"> text </figure><br>$\Longrightarrow$ \begin{figure}[pos] text \end{figure} |
| rule | <rule voffset="off" width="width" height="height"/><br>$\Longrightarrow$ \rule[off]{width}{height} |
| marginpar | <marginpar><left> textleft </left> textright </marginpar><br>$\Longrightarrow$ \marginpar[textleft]{textright} |
| picture<br>Wewnątrz znaczników przechodzimy bezpośrednio do trybu TEX | <picture size="size" offset="offset"><br><?tex text ?></picture><br>$\Longrightarrow$ \begin{picture}(size)(offset) text<br>\end{picture} |
| minipage | <minipage pos="pos" width="width"> text </minipage><br>$\Longrightarrow$ \begin{minipage}[pos]{width} text<br>\end{minipage} |
| framebox, makebox | <makebox pos="pos" width="width"> text </makebox><br>$\Longrightarrow$ \makebox[width][pos]{text} |
| parbox | <parbox pos="pos" width="width"> text </box><br>$\Longrightarrow$ \parbox[pos]{width}{text} |
| addvspace | <addvspace length="length"/><br>$\Longrightarrow$ \addvspace{length} |
| thebibliography | <thebibliography offset="offset"> text </thebibliography><br>$\Longrightarrow$ \begin{thebibliography}{offset} text<br>\end{thebibliography} |
| bibliographystyle | <bibliographystyle name="name"/><br>$\Longrightarrow$ \bibliographystyle{name} |
| bibitem, cite | <bibitem text="text" id="id"/><br>$\Longrightarrow$ \bibitem[text]{id} |
| pagenumbering | <pagenumbering name="name"/><br>$\Longrightarrow$ \pagenumbering{name} |
| shortstack | <shortstack pos="pos"> text </shortstack><br>$\Longrightarrow$ \shortstack[pos]{text} |
| label, pageref, ref | <label id="id"/> $\Longrightarrow$ \label{id} |
| nowa linia | <br star="yes" skip="skip"/> $\Longrightarrow$ *[skip] |
| inne | <TeX/> $\Longrightarrow$ \TeX{}<br><LaTeX/> $\Longrightarrow$ \LaTeX{}<br><cr/> $\Longrightarrow$ \cr<br><italcor/> $\Longrightarrow$ \/<br><ndash/> $\Longrightarrow$ ^^^^2012<br><mdash/> $\Longrightarrow$ ^^^^2014<br><nbsp/> $\Longrightarrow$ ~<br><par/> $\Longrightarrow$ \par<br><spc/> $\Longrightarrow$ \␣<br><tab/> $\Longrightarrow$ & (w środowisku tabular) lub: \>(w środowisku tabbing)<br><tabdef/> $\Longrightarrow$ \= |
| środowiska nie przekształcane na XƎLaTeX | <?tex text ?> $\Longrightarrow$ text<br><?math text ?> $\Longrightarrow$ \$text\$<br><?displaymath text ?> $\Longrightarrow$ \$\$text\$\$<br><?equation text ?> $\Longrightarrow$ \begin{equation}text<br>\end{equation}<br><?verbatim text ?> $\Longrightarrow$ \begin{myverb}text<br>\end{myverb}<br><?verbatimstar text ?> $\Longrightarrow$ \begin{myverb*}text<br>\end{myverb*}<br><?alltt text ?> $\Longrightarrow$ \begin{alltt}text\end{alltt} |