Data Acquisition and Enhancements

Michal Růžička, Petr Sojka

<xruzick7@fi.muni.cz>, <sojka@fi.muni.cz>

28th April 2010

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Born-digital systems

- Main idea: born-digital data acquisition as a side-effect of publishing printed version of the journal.
- Sometimes the complex journal processing system is too complex.
  - Highly interfering with the current workflow of the editor.
  - Not all the editors use (and are ready to use) LaTeX.
  - Not all the editors use (and are ready to use) BibTeX.
- A simple, universal and flexible solution was needed.

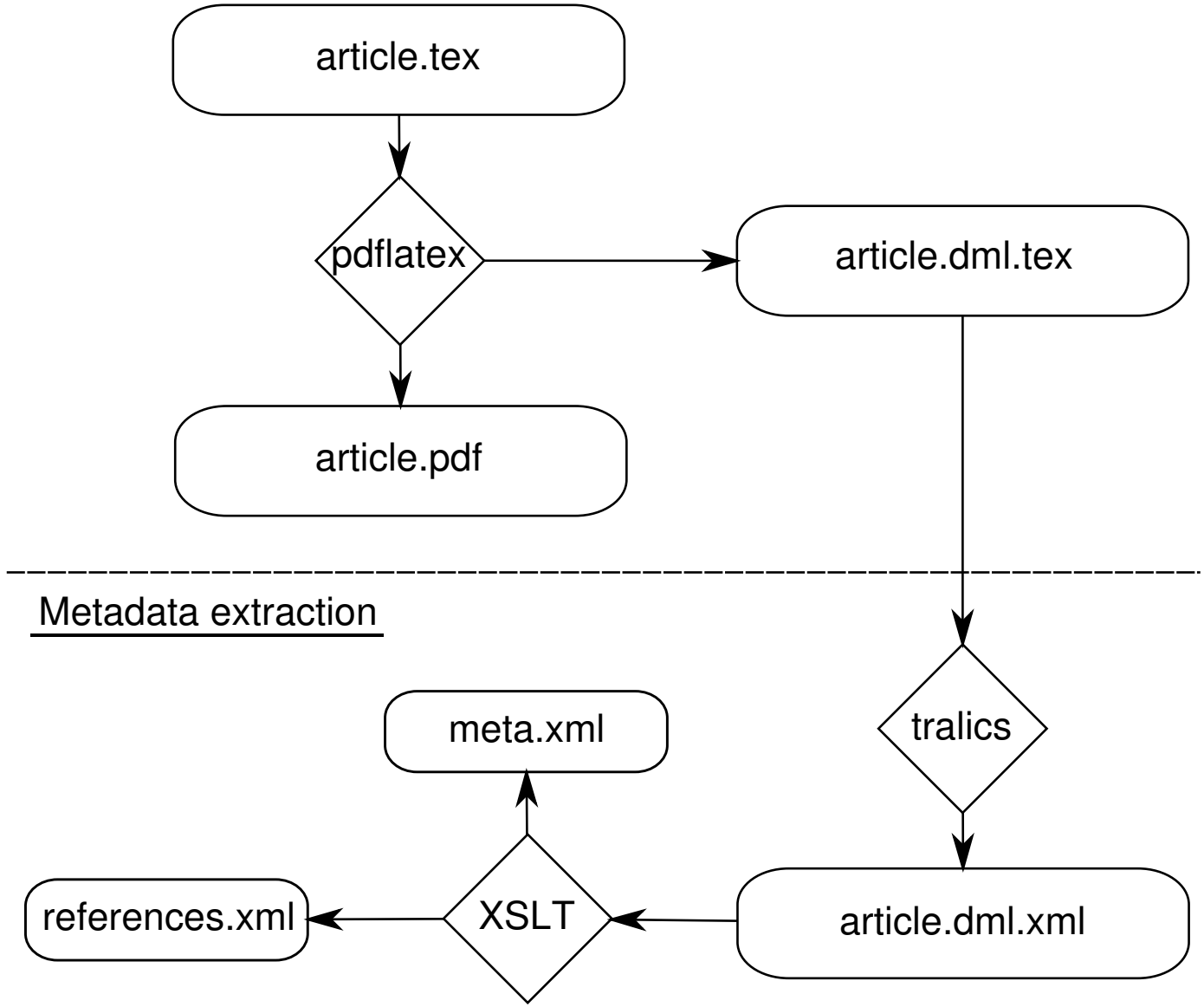
Minimalistic Support of XML Metadata Extraction

- We prepared a minimalistic set of LaTeX macros in the form of LaTeX macro package.
- Macros can be easily customized (if needed) to meet specific needs of a particular editor.
- Macros are simple.
  - The LaTeX macro package itself does not do any transformation of the LaTeX source code to the XML format.
  - Selected parts of the LaTeX document are literally (i.e. without expansion of the TeX code) exported to an external file in such a way that it forms a simple LaTeX document.
- The external file is consequently processed by the Tralics program.

Minimalistic System is Flexible

- This approach is acceptable to all the involved editors
  - Current TEX processing used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No use of BibTEX.
    - A special set of macros is used to mark up the structure of bibliography records.
  - MathML is supported by Tralics.

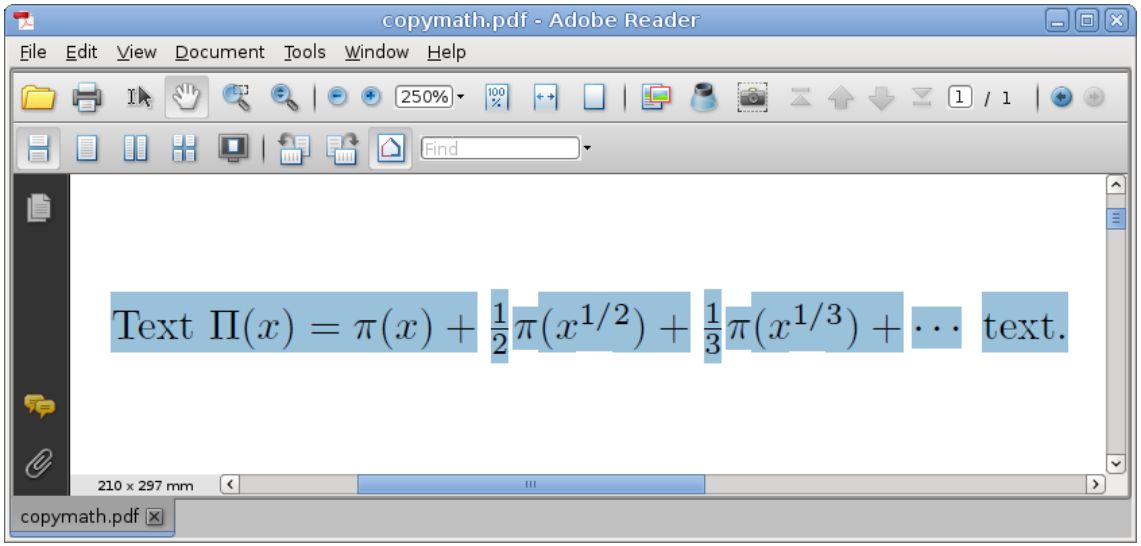
Article processing



Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
- Thanks to pdfTEX, PDF is de facto standard output format of the modern TEX distributions.
- TEX mathematical notation is well known and effective.
  - Also used in other system, e.g. Wikipedia.
- TEX source code is usually good choice for plain text representation of a mathematical expressions.

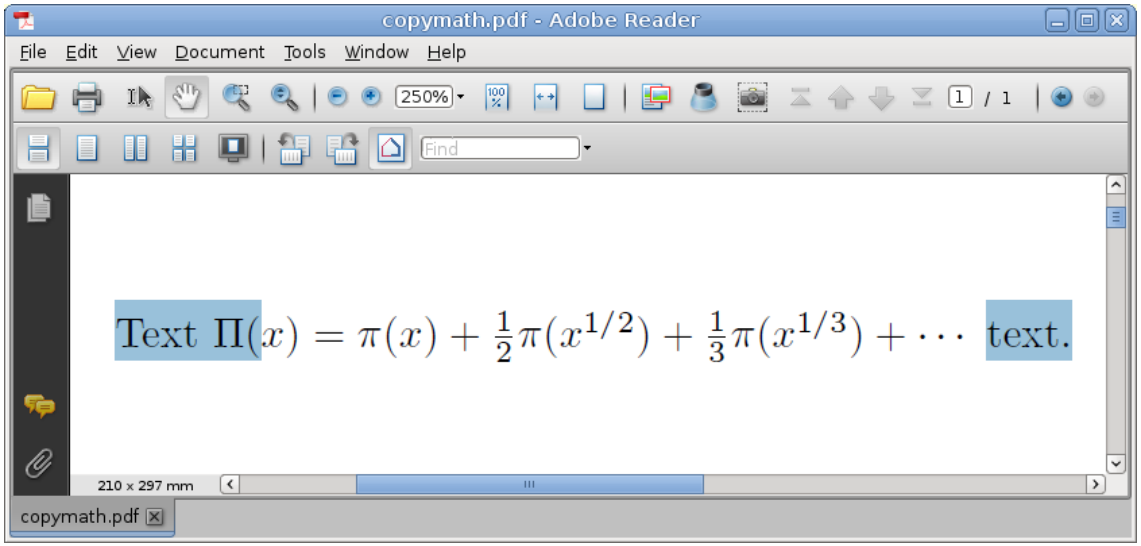
Standard PDF document



Copied text:

```
Text   ( ) =   ( ) + 1
2 ( 1/2) + 1
3 ( 1/3) + · · · text.
```

Copy-Math-enabled PDF document



Copied text:

```
Text $\Pi (x) = \pi (x) +
      \frac {1}{2}\pi (x^{1/2}) +
      \frac {1}{3}\pi (x^{1/3}) + \cdots $
text.
```

Implementation

- Implemented using the `ActualText` command of the PDF language.
- Requires quite a nonstandard modifications of the TEX mathematical environments.

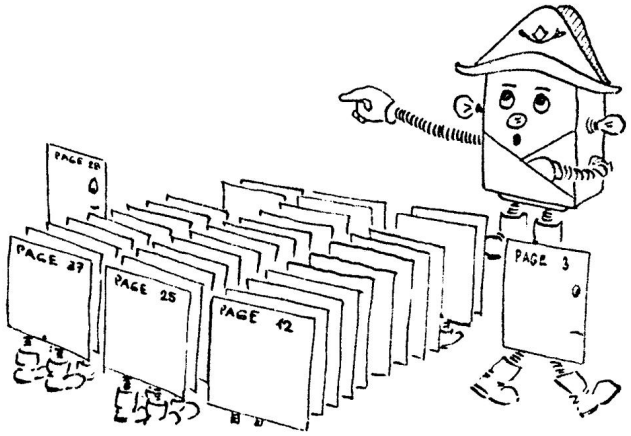
Implementation

```
%% Auxiliary macros.
\newcounter{nestedmath} \setcounter{nestedmath}{0}
%
\newtoks\copymath@envgetbuffera
\newtoks\copymath@envgetbufferb
%
\long\def\copymath@envget#1#2\end #3{%
    \copymath@envgetbuffera=\expandafter{\copymathenvput}%
    \def\copymath@envtempa{#3}\def\copymath@envtempb{#1}%
        \ifx\copymath@envtempa\copymath@envtempb%
            \copymath@envgetbufferb={#2}%
            \def\copymath@envgetnext{\end{#1}}%
    \else%
        \copymath@envgetbufferb={#2\end{#3}}%
        \def\copymath@envgetnext{\copymath@envget{#1}}%
    \fi%
    \global\edef\copymathenvput{%
        \the\copymath@envgetbuffera \the\copymath@envgetbufferb}%
    \copymath@envgetnext}
%
\long\def\copymathenvget#1{%
    \gdef\copymathenvput{}\copymath@envget{#1}}
%

%% $
\let\@origensuredmath=\@ensuredmath

\def\normalinlinemath#1{%
 \ifnum\value{nestedmath}>0 \@origensuredmath{#1}%
 \else%
    \addtocounter{nestedmath}{1}%
```

Implementation

- Implemented using the `ActualText` command of the PDF language.
- Requires quite a nonstandard modifications of the TEX mathematical environments.
- Still experimental.

Questions?



References

Apics Team.

Tralics: a LaTeX to XML translator [online].
Last modified $Date: 2009/11/24 17:17:03 $ [cit. 2010-02-10].
Dostupný z WWW: <http://www-sop.inria.fr/apics/tralics/>.

Bouche, Thierry.

CEDRICS: When CEDRAM Meets Tralics.
In: Sojka Petr (editor): DML 2008 – Towards Digital Mathematics Library, Birmingham, UK, July 27th, 2008, 153–165.

Růžička, Michal.

A Journal Processing System [online].
[cit. 2010-03-16].
Available from WWW: <http://www.fi.muni.cz/lemma/math-journal/>.