Data Enhancements in a Digital Mathematical Library

Michal Růžička and Petr Sojka

Masaryk University, Faculty of Informatics
Brno, Czech Republic
<mruzicka@mail.muni.cz>, <sojka@fi.muni.cz>

July 7th, 2010

EuDML

The EUROPEAN DIGITAL
MATHEMATICS LIBRARY

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library

- The quality of digital mathematical library depends on the quality of data it offers.
- The viability of a digital library rests with new acquisitions emerging mainly in the form of born-digital publications.
- It is important to
  - provide data as soon as possible,
  - in a digital-use-friendly format,
  - and exactly matching printed originals.

Goals of a Digital Library (cont.)

- In this talk we are going to show
  - a lightweight XML metadata extraction system for mathematical journal editors,
  - a proof of concept of a method that improves usability of mathematical PDF documents.

Goals of a Digital Library (cont.)

- In this talk we are going to show
  - a lightweight XML metadata extraction system for mathematical journal editors,
  - a proof of concept of a method that improves usability of mathematical PDF documents.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
    - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

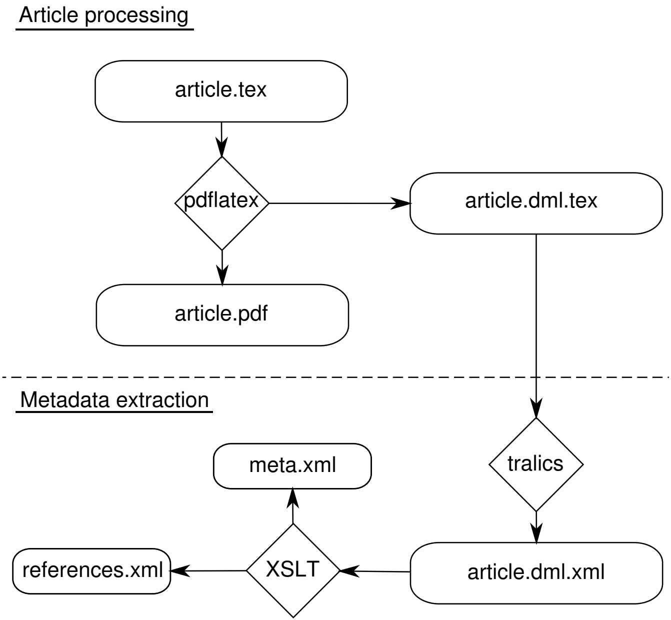
- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
    - Great interference 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.

Need for a Lightweight XML Metadata Extraction System

- It has been necessary to prepare appropriate software support for the mathematical journals involved in the DML-CZ project that will enable editors to prepare born-digital data easily.
- Main idea: born-digital data acquisition as a by-product of publishing printed version of the journal.
- The first approach was a complex system inspired by the French CEDRAM project.
- Sometimes the complex journal processing system is too complex.
  - Great interference 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.


Article processing
article.tex
pdflatex
article.dml.tex
article.pdf
Metadata extraction
meta.xml
tralics
references.xml
XSLT
article.dml.xml

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
    - Can be easily customized to meet needs of a particular journal document class / style file.
    - The LaTeX macro package itself does not transform the LaTeX source code to XML.
    - Literally exports selected parts of the LaTeX document to an external file.
    - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
    - Can be easily customized to meet needs of a particular journal document class / style file.
    - The LaTeX macro package itself does not transform the LaTeX source code to XML.
    - Literally exports selected parts of the LaTeX document to an external file.
    - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
  - Can be easily customized to meet needs of a particular journal document class / style file.
  - The LaTeX macro package itself does not transform the LaTeX source code to XML.
  - Literally exports selected parts of the LaTeX document to an external file.
  - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
  - Can be easily customized to meet needs of a particular journal document class / style file.
  - The LaTeX macro package itself does not transform the LaTeX source code to XML.
  - Literally exports selected parts of the LaTeX document to an external file.
  - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
    - Can be easily customized to meet needs of a particular journal document class / style file.
    - The LaTeX macro package itself does not transform the LaTeX source code to XML.
    - Literally exports selected parts of the LaTeX document to an external file.
    - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work (cont.)

```
\documentclass[runningheads]{llncs}
\usepackage{dmlcommon}
\usepackage{dmlcz}

\begin{document}

\author{Petr Sojka}
\dmlaindex{Sojka}{Petr}
\dmltitle{Towards a Digital Mathematical Library}
...
\maketitle

\begin{dmlabstract}
The workshop's objectives were to formulate the strategy
and goals of a global mathematical digital library...
\end{dmlabstract}
...
```

How Does It Work (cont.)

```
\documentclass{dmlczmeta}\begin{document}

\begin{xmlelement}{author}{Sojka, Petr
\XMLaddatt{order}{1}}\end{xmlelement}

\begin{xmlelement}{title}{Towards a Digital Mathematical
Library\XMLaddatt{lang}{eng}}\end{xmlelement}

\begin{xmlelement}{abstract}\XMLaddatt{lang}{eng}\bgroup
The workshop's objectives were to formulate the strategy
and goals of a global mathematical digital library...
\egroup\end{xmlelement}

\begin{xmlelement}{keyword}{OCR\XMLaddatt{lang}{eng}}
\end{xmlelement}

...
\end{document}
```

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
    - The most indispensable part of the system.
    - Its engine is able to process regular LaTeX code.
    - It is not necessary to
        - convert the LaTeX code to plain text directly,
        - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
    - The most indispensable part of the system.
    - Its engine is able to process regular LaTeX code.
    - It is not necessary to
        - convert the LaTeX code to plain text directly,
        - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE std SYSTEM 'classes.dtd'>
<!-- Translated from latex by tralics 2.13.5,
     date: 2010/07/03-->
<std><p>
<author order='1'>Sojka, Petr</author>
<title lang='eng'>Towards a Digital Mathematical
Library</title>

<abstract lang='eng'>The workshop's objectives were to
formulate the strategy...</abstract>
<keyword lang='eng'>OCR</keyword>
<keyword lang='eng'>OpenMath</keyword>

<language>eng</language>
<abstractlanguage>eng</abstractlanguage>
...
</p></std>
```

How Does It Work (cont.)

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
  <title lang="eng">Towards a Digital
    Mathematical Library</title>

  <author order="1">Sojka, Petr</author>

  <language>eng</language>

  <keyword lang="eng">OCR</keyword>
  <keyword lang="eng">OpenMath</keyword>

  <summary lang="eng">The workshop's objectives
    were to formulate the strategy...</summary>
  ...
</article>
```

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
    - Current TEX processing is used.
    - Platform independent.
        - The TEX itself produces the source file.
        - XML generated using Tralics and XSLT.
    - No need for BibTEX.
- It is safe.
    - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
    - Current TEX processing is used.
    - Platform independent.
        - The TEX itself produces the source file.
        - XML generated using Tralics and XSLT.
    - No need for BibTEX.
- It is safe.
    - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

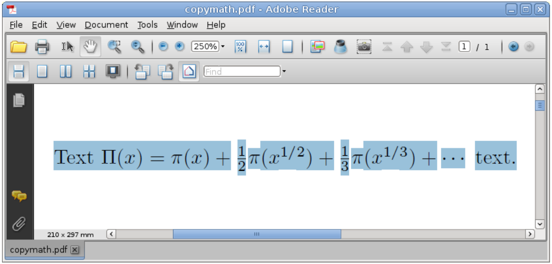
Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
    - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
    - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Standard PDF document

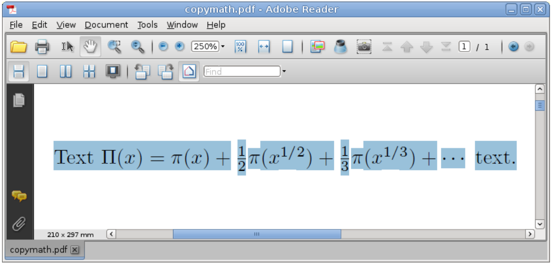

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

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

LaTeX source code:

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

Standard PDF document

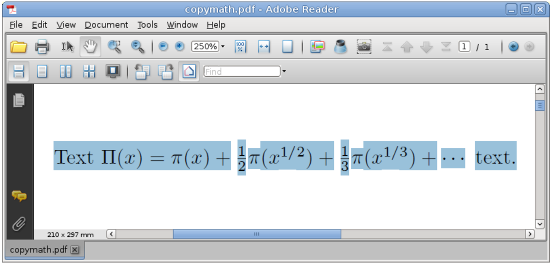

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

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

PDF code:

```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ/F17 9.9626 Tf 23.247 0 Td
[(\005\050)]TJ/F20 9.9626 Tf 11.346 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[(\051)-278(=)]TJ/F20 9.9626 Tf 17.158 0 Td [(\031)]TJ/F17 9.9626 Tf 6.036 0 Td
[(\050)]TJ/F20 9.9626 Tf 3.875 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923 Td [(1)]TJ
ET
```

Standard PDF document

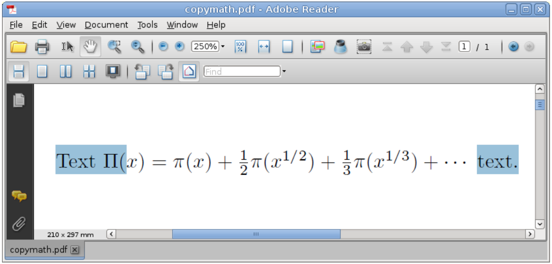

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

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

Text obtained using Copy & Paste function of PDF reader:

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

CopyMath-enabled PDF document

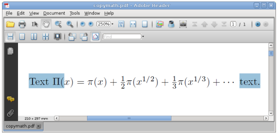

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
Find
210 x 297 mm
copymath.pdf

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

LaTeX source code:

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

CopyMath-enabled PDF document

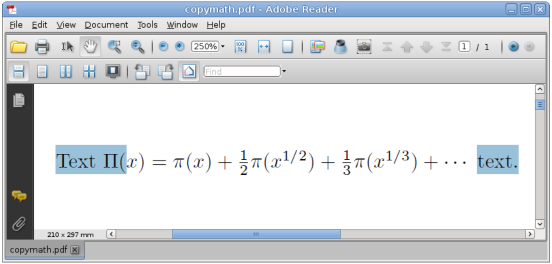

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

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

PDF code:

```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ
ET
1 0 0 1 171.959 707.125 cm
/Span <<
/ActualText<245C506920287829203D205C706920287829202B205C66726163207B317D7B32
7D5C70692028785E7B312F327D29202B205C66726163207B317D7B337D5C70692028785E7B31
2F337D29202B205C63646F74732024> >> BDC
1 0 0 1 -171.959 -707.125 cm
BT
/F17 9.9626 Tf 171.959 707.125 Td [(\005\050)]TJ/F20 9.9626 Tf 11.346 0 Td
[(x)]TJ/F17 9.9626 Tf 5.694 0 Td [(\051)-278(=)]TJ/F20 9.9626 Tf 17.158 0 Td
[(\031)]TJ/F17 9.9626 Tf 6.036 0 Td [(\050)]TJ/F20 9.9626 Tf 3.875 0 Td
[(x)]TJ/F17 9.9626 Tf 5.694 0 Td [(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923
Td [(1)]TJ
ET
```

CopyMath-enabled PDF document


copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

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

Text obtained using Copy & Paste function of PDF reader:

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

Implementation

- The `ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not easy.
  - This requires nonstandard modifications of the LaTeX mathematical environments.

Implementation

- The `ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not easy.
  - This requires nonstandard modifications of the LaTeX mathematical environments.

Implementation

- The `ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not easy.
  - This requires nonstandard modifications of the LaTeX mathematical environments.

Implementation

- The `ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not easy.
  - This requires nonstandard modifications of the LaTeX mathematical environments.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- Still experimental.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- Still experimental.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- Still experimental.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- Still experimental.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- Still experimental.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
  - It is used by journals
    - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
    - Acta Mathematica Universitatis Ostraviensis,
    - Archivum Mathematicum,
    - Kybernetika,
    - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
  - It is used by journals
    - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
    - Acta Mathematica Universitatis Ostraviensis,
    - Archivum Mathematicum,
    - Kybernetika,
    - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
  - It is used by journals
    - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
    - Acta Mathematica Universitatis Ostraviensis,
    - Archivum Mathematicum,
    - Kybernetika,
    - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
    - It is used by journals
        - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
        - Acta Mathematica Universitatis Ostraviensis,
        - Archivum Mathematicum,
        - Kybernetika,
        - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
    - It is used by journals
        - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
        - Acta Mathematica Universitatis Ostraviensis,
        - Archivum Mathematicum,
        - Kybernetika,
        - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

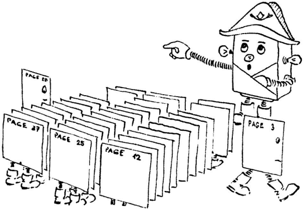
Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
  - It is used by journals
    - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
    - Acta Mathematica Universitatis Ostraviensis,
    - Archivum Mathematicum,
    - Kybernetika,
    - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Conclusions

- Minimalist modifications of the current editorial workflow proved to be an easy way of moving mathematical journal editors to a digital-library-friendly state.
    - It is used by journals
        - Acta Universitatis Palackianae Olomucensis (Facultas Rerum Naturalium, Mathematica),
        - Acta Mathematica Universitatis Ostraviensis,
        - Archivum Mathematicum,
        - Kybernetika,
        - Proceedings of DML workshop.
- The CopyMath macro package shows an alternative route to improving pdfTEX-generated PDFs, but the proper redefinition of all possible mathematical environments cannot expected to be easy.

Questions?



Czech Digital Mathematics Library [online].
[cit. 2010-04-24].
Available from WWW: <http://dml.cz/>.

EuDML: The European Digital Mathematics Library [online].
This page was last modified on 20 January 2010, at 08:09. [cit. 2010-04-25].
Available from WWW: <http://www.eudml.eu/>.

Bouche, T.:
A pdfLaTeX-based automated journal production system.
In Proceedings of EuroTeX 2006, TUGboat **27**(1) (2006) 45–50.

Centre de diffusion de revues académiques mathématiques [Center for diffusion of mathematic journals] [online].
[cit. 2008-05-25].
Available from WWW: <http://www.cedram.org/>.

Růžička, M.:
Automated Processing of TeX-Typeset Articles for a Digital Library.
In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, 167–176.

Archivum Mathematicum [online].
Masaryk University, Brno, Czech Republic.
Last modified December 18, 2009 [cit. 2010-04-25].
Available from WWW: <http://www.emis.de/journals/AM/>.

Grimm, J.:
Tralics, a LaTeX to XML Translator.
In Proceedings of EuroTeX, TUGboat **24**(3) (2003) 377–388.

Tralics: a LaTeX to XML translator [online].
Last modified $Date: 2009/11/24 17:17:03 $ [cit. 2010-04-24].
Available from WWW: <http://www-sop.inria.fr/apics/tralics/>.

Infty Project: Research Project on Mathematical Information Processing [online].

[cit. 2010-06-02].
Available from WWW: <http://www.inftyproject.org/en/>.

Suzuki, M.; Kanahori, T.; Ohtake, N.; Yamaguchi, K.:
An Integrated OCR Software for mathematical Documents and Its Output with Accessibility.
Computers Helping people with Special Needs, 9th International Conference ICCHP2004, Paris, July 2004, Lecture Notes in Computer Sciences 3119, Springer (2004) 648–655.