[Meta]data acquisition and validation

Michal Růžička

<xruzick7@fi.muni.cz>

17th March 2010

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Retro-born-digital Journals

- Documents are available in a digital format.
- Mathematical journals ⇒ massive use of $\TeX$.
- Different formats, not suitable for direct import into a digital library.
- Recompilation of all documents was impossible.
  - Missing source files.
  - Missing older versions of appropriate $\TeX$-style files.

Retro-born-digital Journals

- Documents are available in a digital format.
- Mathematical journals ⇒ massive use of TEX.
- Different formats, not suitable for direct import into a digital library.
- Recompilation of all documents was impossible.
  - Missing source files.
  - Missing older versions of appropriate TEX-style files.

Retro-born-digital Journals

- Documents are available in a digital format.
- Mathematical journals ⇒ massive use of TEX.
- Different formats, not suitable for direct import into a digital library.
- Recompilation of all documents was impossible.
  - Missing source files.
  - Missing older versions of appropriate TEX-style files.

Retro-born-digital Journals

- Documents are available in a digital format.
- Mathematical journals ⇒ massive use of TEX.
- Different formats, not suitable for direct import into a digital library.
- Recompilation of all documents was impossible.
  - Missing source files.
  - Missing older versions of appropriate TEX-style files.

Conversion of Fulltexts

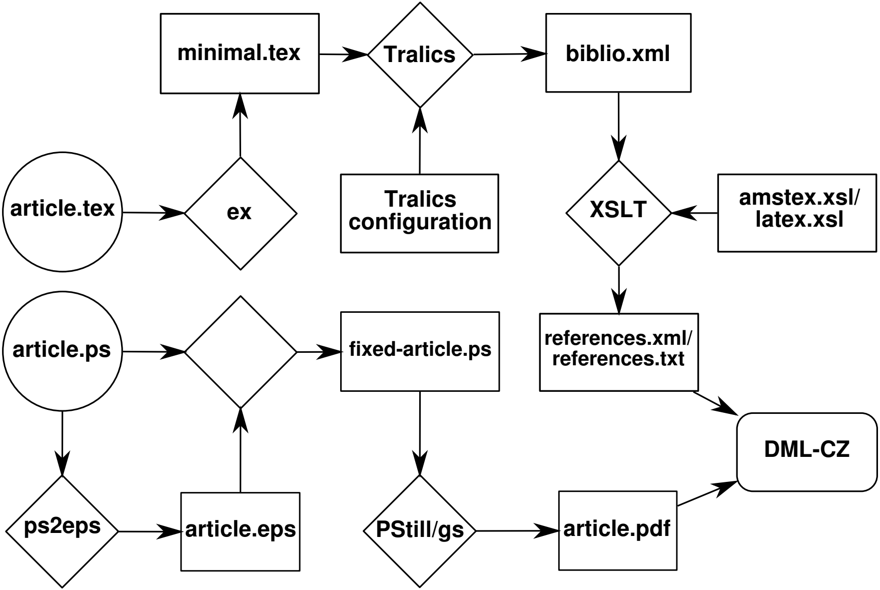
- Original PostScript files contained 300 DPI bitmap fonts.
- Attempt to replace the bitmap fonts with their outline alternatives.
    - The FixFont program failed.
    - Partial success with the PStill program.
        - PStill depends on `dvips` comment lines within the PostScript file.

Conversion of Fulltexts

- Original PostScript files contained 300 DPI bitmap fonts.
- Attempt to replace the bitmap fonts with their outline alternatives.
    - The FixFont program failed.
    - Partial success with the PStill program.
        - PStill depends on `dvips` comment lines within the PostScript file.

Conversion of Fulltexts

- Original PostScript files contained 300 DPI bitmap fonts.
- Attempt to replace the bitmap fonts with their outline alternatives.
  - The FixFont program failed.
  - Partial success with the PStill program.
    - PStill depends on `dvips` comment lines within the PostScript file.


minimal.tex
Tralics
biblio.xml
article.tex
ex
Tralics configuration
XSLT
amstex.xsl/ latex.xsl
article.ps
fixed-article.ps
references.xml/ references.txt
DML-CZ
ps2eps
article.eps
PStill/gs
article.pdf

Born-digital Journals

- Main idea: born-digital data acquisition as a side-effect of publishing printed version of the journal.
- With the aim of automating as much as possible…
  - Complex journal processing system.
- …do not impair proven workflow of the editorial staff, however.
  - Minimalistic journal processing system.

Born-digital Journals

- Main idea: born-digital data acquisition as a side-effect of publishing printed version of the journal.
- With the aim of automating as much as possible…
  - Complex journal processing system.
- …do not impair proven workflow of the editorial staff, however.
  - Minimalistic journal processing system.

Born-digital Journals

- Main idea: born-digital data acquisition as a side-effect of publishing printed version of the journal.
- With the aim of automating as much as possible…
  - Complex journal processing system.
- …do not impair proven workflow of the editorial staff, however.
  - Minimalistic journal processing system.

Tralics

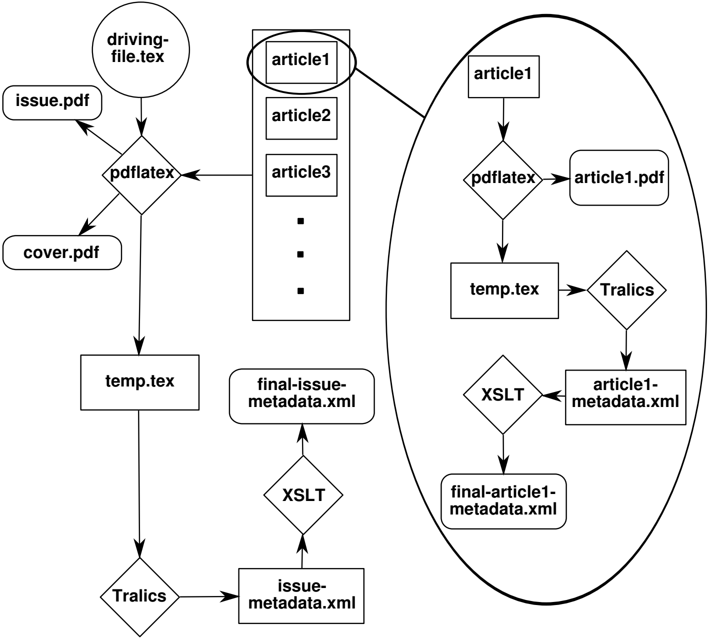
- A LaTeX to XML translator.
- Free software, available for different platforms.
  - GNU/Linux, Apple MacOS X, Microsoft Windows.
- Able to deal directly with a LaTeX source file and BibTeX database.
  - Results of the conversion are quite good.
- Used for both retro-born- and born-digital system.

Tralics

- A LaTeX to XML translator.
- Free software, available for different platforms.
  - GNU/Linux, Apple MacOS X, Microsoft Windows.
- Able to deal directly with a LaTeX source file and BibTeX database.
  - Results of the conversion are quite good.
- Used for both retro-born- and born-digital system.

Tralics

- A LaTeX to XML translator.
- Free software, available for different platforms.
  - GNU/Linux, Apple MacOS X, Microsoft Windows.
- Able to deal directly with a LaTeX source file and BibTeX database.
  - Results of the conversion are quite good.
- Used for both retro-born- and born-digital system.


driving-file.tex
issue.pdf
pdflatex
cover.pdf
temp.tex
Tralics
issue-metadata.xml
XSLT
final-issue-metadata.xml
article1
article2
article3
article1
pdflatex
article1.pdf
temp.tex
Tralics
article1-metadata.xml
XSLT
final-article1-metadata.xml

Metadata Validation

- Editorial staff generated metadata ⇒ need of validation.
  - It is not easy to find reasonably strict schema.
- TEX code must be correct.
  - The code is compiled during generation of the digital-library-specific cover of the article.
  - Conversion to MathML.
- On-line application.
  - Integrates validation and delivery of the final data to the digital library.
  - Final validation is mandatory.

Metadata Validation

- Editorial staff generated metadata ⇒ need of validation.
  - It is not easy to find reasonably strict schema.
- TEX code must be correct.
  - The code is compiled during generation of the digital-library-specific cover of the article.
  - Conversion to MathML.
- On-line application.
  - Integrates validation and delivery of the final data to the digital library.
  - Final validation is mandatory.

Metadata Validation

- Editorial staff generated metadata ⇒ need of validation.
  - It is not easy to find reasonably strict schema.
- TEX code must be correct.
  - The code is compiled during generation of the digital-library-specific cover of the article.
  - Conversion to MathML.
- On-line application.
  - Integrates validation and delivery of the final data to the digital library.
  - Final validation is mandatory.

Data Security

- EuDML will integrate national digital libraries across Europe.
  - Different sources of data with different access policies.
  - A lot of people have to be able to access data.
- Some parts of the repository are not public (moving walls etc.), restricted by the license of the provider.
  - Sometimes it is necessary to access all the articles – finding similar articles.
- The YADDA platform used in the EuDML possesses its own powerful Authentication and Authorization Service (YADDA-AAS).
- DML-CZ uses digital signatures to mark each article accessible from the public repository.

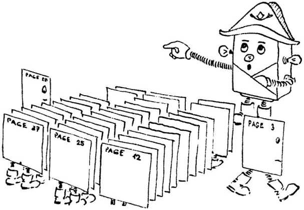
Data Security

- EuDML will integrate national digital libraries across Europe.
  - Different sources of data with different access policies.
  - A lot of people have to be able to access data.
- Some parts of the repository are not public (moving walls etc.), restricted by the license of the provider.
  - Sometimes it is necessary to access all the articles – finding similar articles.
- The YADDA platform used in the EuDML possesses its own powerful Authentication and Authorization Service (YADDA-AAS).
- DML-CZ uses digital signatures to mark each article accessible from the public repository.

Data Security

- EuDML will integrate national digital libraries across Europe.
  - Different sources of data with different access policies.
  - A lot of people have to be able to access data.
- Some parts of the repository are not public (moving walls etc.), restricted by the license of the provider.
  - Sometimes it is necessary to access all the articles – finding similar articles.
- The YADDA platform used in the EuDML possesses its own powerful Authentication and Authorization Service (YADDA-AAS).
- DML-CZ uses digital signatures to mark each article accessible from the public repository.

Questions?



References

Probets, S., Brailsford, D.:
Substituting outline fonts for bitmap fonts in archived PDF files.
Software-Practice and Experience. **33**(9) (2003) 885–899.
ISSN: 0038-0644.

Research - Fonts [online].
[cit. 2010-02-10].
Available from WWW: <http://www.eprg.org/research/fonts/>.

Siegert, F.:
PStill: ...generate, reprocess, normalize and extract content for PDF, EPS and PS [online].
[cit. 2010-02-10].
Available from WWW: <http://www.pstill.com/>.

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/>.