Document Engineering for a Digital Library

PDF recompression using JBIG2 and other optimization of PDF

Petr Sojka et al.

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

## University of Portsmouth Computing Seminar, Portsmouth, UK

November 5th, 2010, 12AM, Anglesea building, room 1.05

Outline and two take-home messages

1. Pictorial overview

2. Motivation, vision of PubMed Central for Mathematics

3. Complexity of digitization workflow of The Czech Digital Mathematics Library DML-CZ

4. Document engineering technologies and tools for DML-CZ and EuDML

5. Tools developed (PDF recompressor et al.)

6. Results: already compressed 2-layer bitonal PDFs squeezed to 38%

7. Summary, conclusions and future work
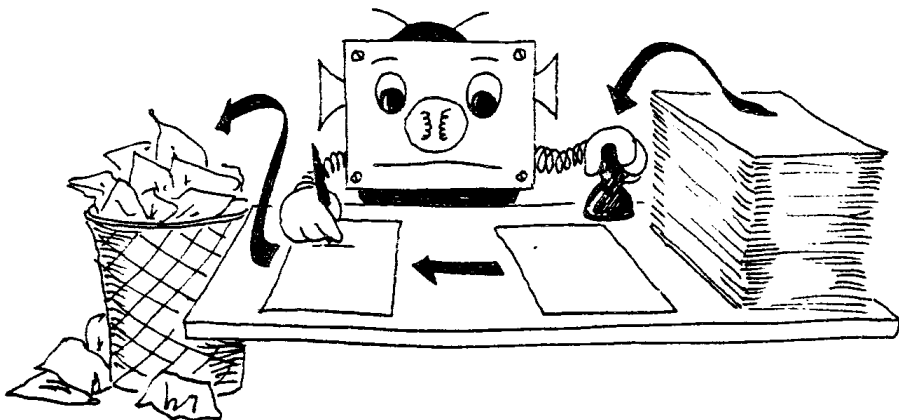
From paper to *digital* library and processing

Information overload in globalized scientific world
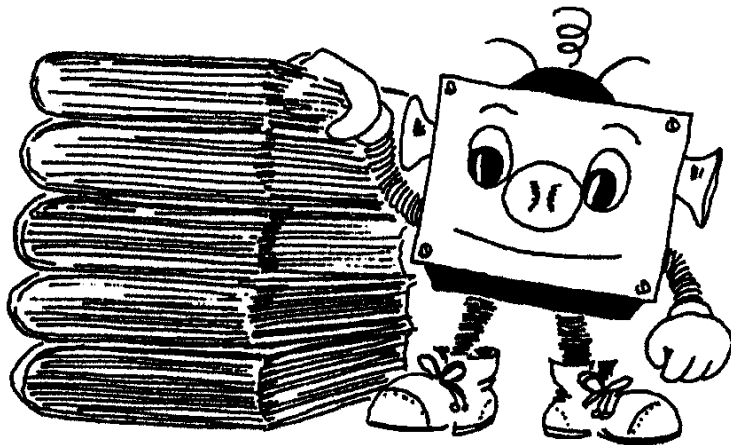
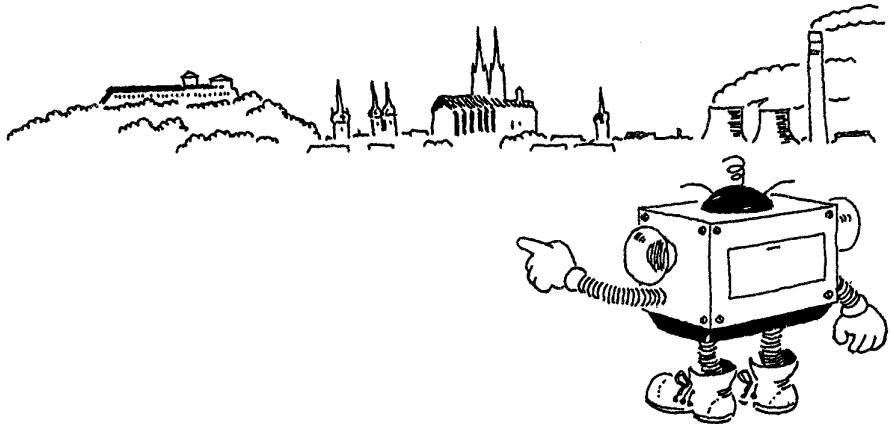# Information overload also in specific domains (mathematics)
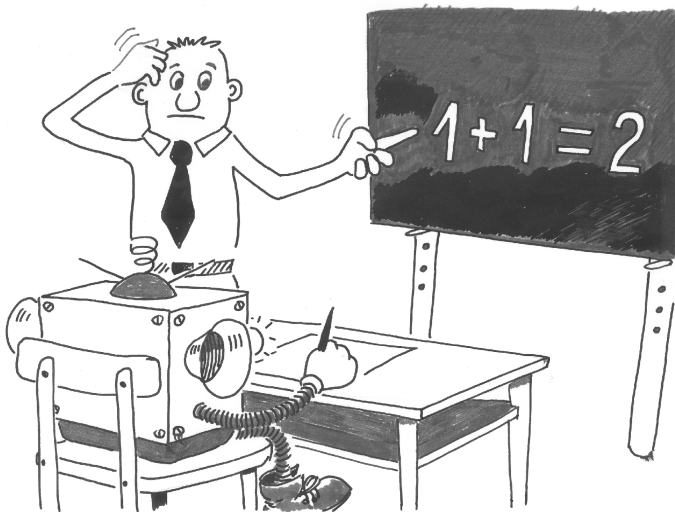
Document Engineering (DocEng): from paper to digital workflow

DocEng: retro-digitization, digital library development

DocEng for specific/local (Brno, CZ) purposes

DocEng in DML-CZ: new workflows and data processing

DocEng in DML-CZ: new tools
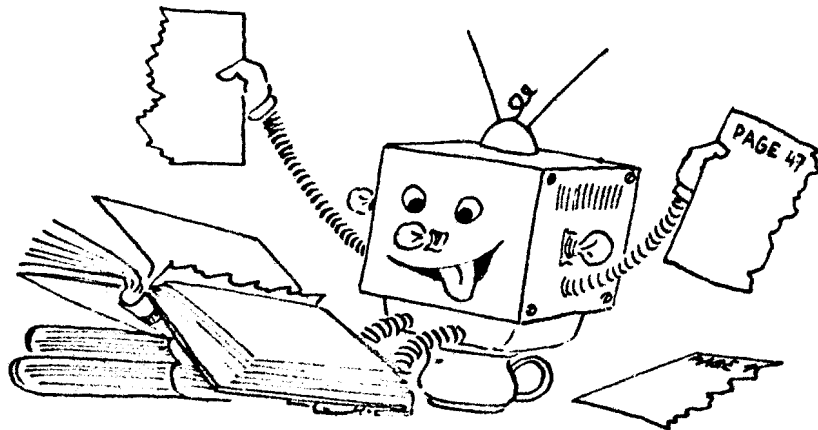
'Bottom up' deployment towards EU or worldwide scale
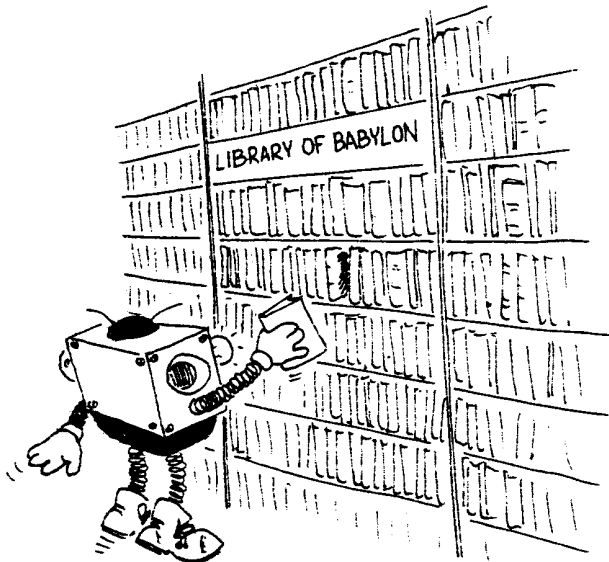
The European Digital Mathematics Library: EuDML

EuDML: from local data collections to the virtual DL
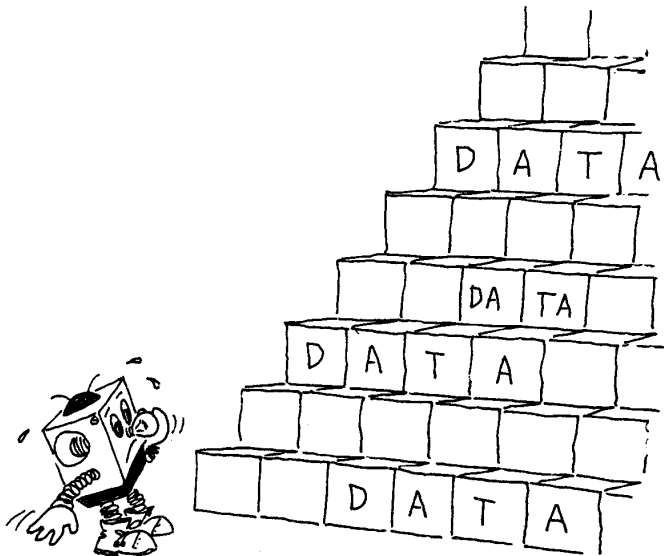
DocEng for EuDML: scalable [PDF] tools development
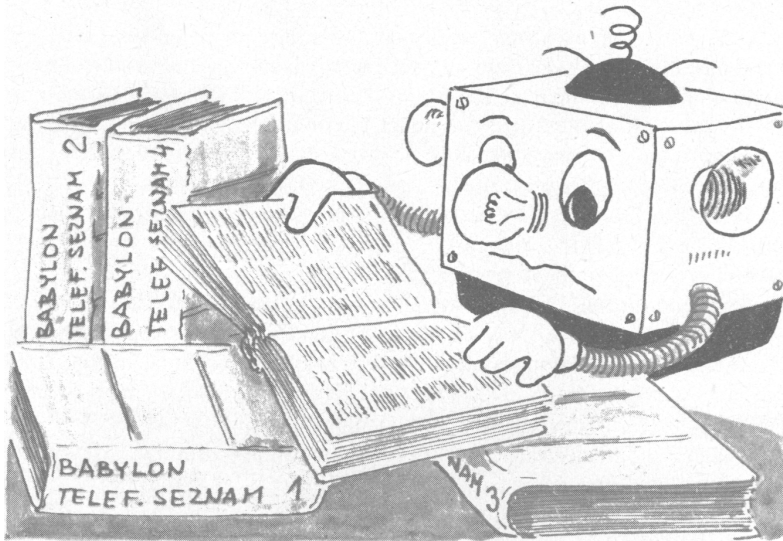
Yes, you can! You can have visibility, scalability, similarity fulltext metrics, 38% of original size PDFs,...

TeX
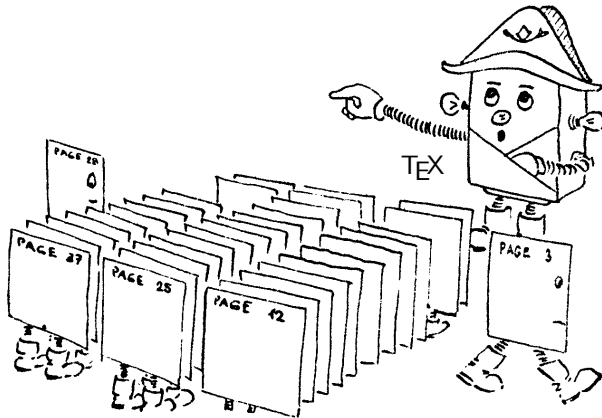
End of talk overview

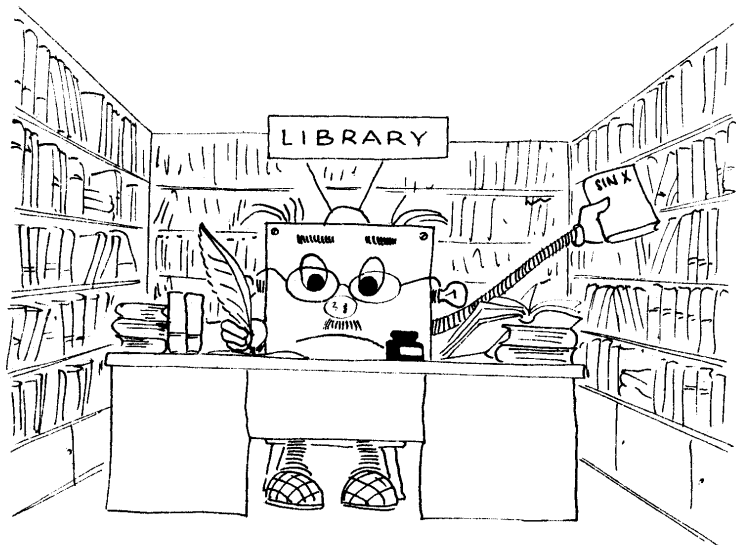# Decade of the vision of WDML as PubMed 4 Math

In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Progress of IT, connectivity, cheap storage, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation ($100,000,000 requested). Application was *not* successful.

Decade of the vision of WDML as PubMed 4 Math

In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Progress of IT, connectivity, cheap storage, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation ($100,000,000 requested). Application was *not* successful.

Decade of the vision of WDML as PubMed 4 Math

In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Progress of IT, connectivity, cheap storage, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation ($100,000,000 requested). Application was *not* successful.

Decade of the vision of WDML as PubMed 4 Math

In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Progress of IT, connectivity, cheap storage, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation ($100,000,000 requested). Application was *not* successful.

Vision of European Digital Mathematics Library

Even other attempts on the European level (FP5, FP6) were not successful. Finally three year project or *European Digital Mathematics Library, EuDML* (programme EU CIP-ICT-PSP, type Pilot B, EU contribution *(1.6 MEur, 50% of total budget only)*

$\mathcal{E}u$DML

The **EUROPEAN DIGITAL**
MATHEMATICS LIBRARY

started from February 2010. The strategy of EuDML is:

- to master the technology, develop tools and offer them;

- concept of *moving wall* to motivate and engage commercial publishers without Open Access bussiness model;

- to collect data (from existing local or publisher's) *digital libraries* into 'one-stop shop' and achieve critical mass in the domain $\rightarrow$ 'a must/me too' effect then as with PubMed.

Vision of European Digital Mathematics Library

Even other attempts on the European level (FP5, FP6) were not successful. Finally three year project or *European Digital Mathematics Library, EuDML* (programme EU CIP-ICT-PSP, type Pilot B, EU contribution *(1.6 MEur, 50% of total budget only)*

started from February 2010. The strategy of is:

- to master the technology, develop tools and offer them;
- concept of *moving wall* to motivate and engage commercial publishers without Open Access bussiness model;
- to collect data (from existing local or publisher's) *digital libraries* into 'one-stop shop' and achieve critical mass in the domain → 'a must/me too' effect then as with PubMed.

EuDML as a virtual library portal

EuDML will be a *virtual* library based on data from smaller data providers, DLs and publishers:
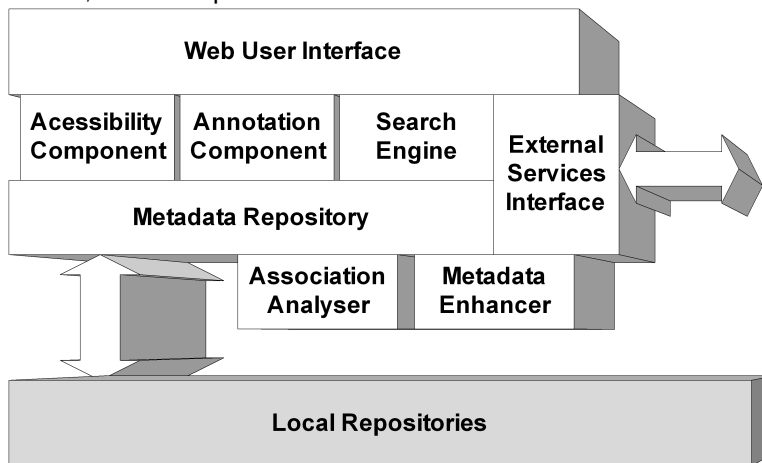
European Digital Mathematics Library
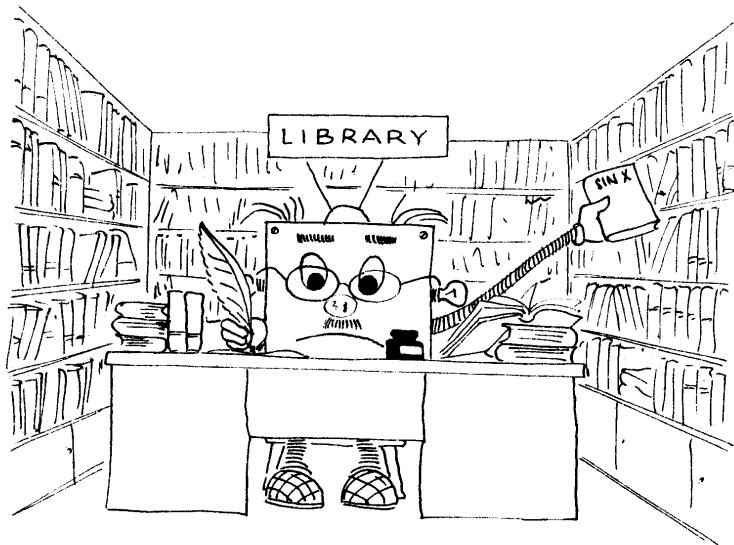
Bottom up—from building bricks of regional repositories

As DML content providers serve mostly publisher's or regional DML repositories as The Czech Digital Mathematics Library DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,...: aggregating content from local repositories to build the bigger (global?) DML.

Example of DML-CZ: up and running digital mathematics library <http://dml.cz> with nearly 30,000 papers (300,000 pages). For more, see (who, what, browse, browse similar, how to search).

Bottom up—from building bricks of regional repositories

As DML content providers serve mostly publisher's or regional DML repositories as The Czech Digital Mathematics Library DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,. . . : aggregating content from local repositories to build the bigger (global?) DML.

Example of DML-CZ: up and running digital mathematics library <http://dml.cz> with nearly 30,000 papers (300,000 pages). For more, see (who, what, browse, browse similar, how to search).

From paper to digital processing, from local to the whole
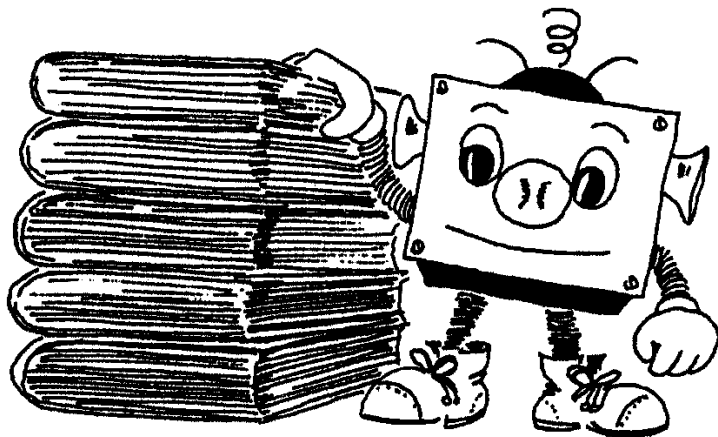
DML-CZ workflow
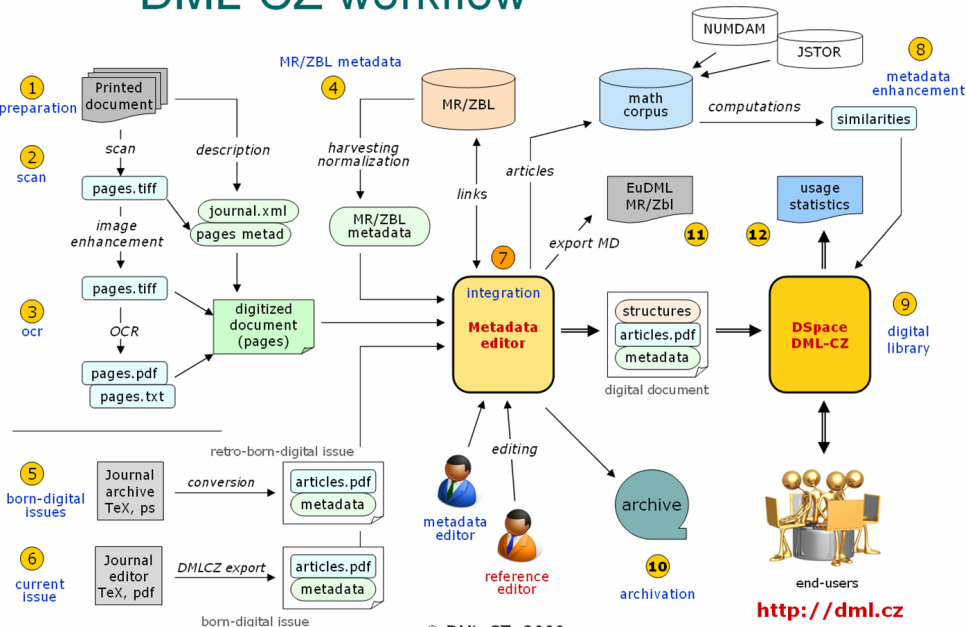
Take care! "God is in the details." (Mies van der Rohe)

Challenges of Math handling: OCR, indexing, search...
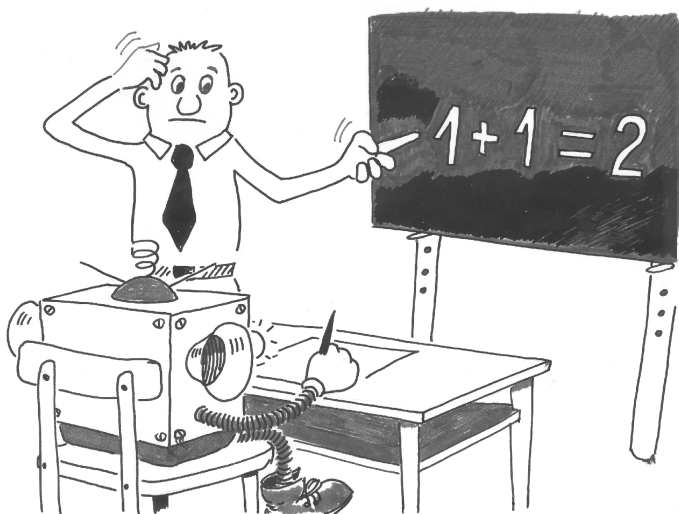
DML-CZ—data: scientific math published in CZ/SK

ИОСИФ ВИССАРИОНОВИЧ СТАЛИН
1879—1953

Document engineering—from paper to digital *workflow*
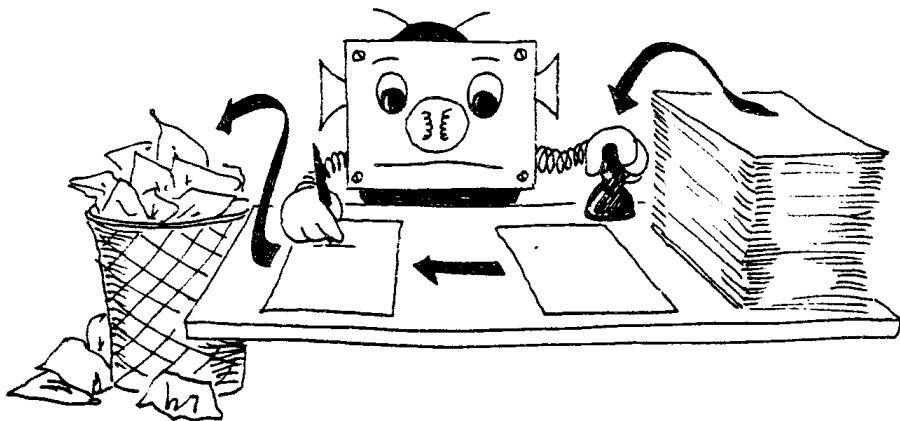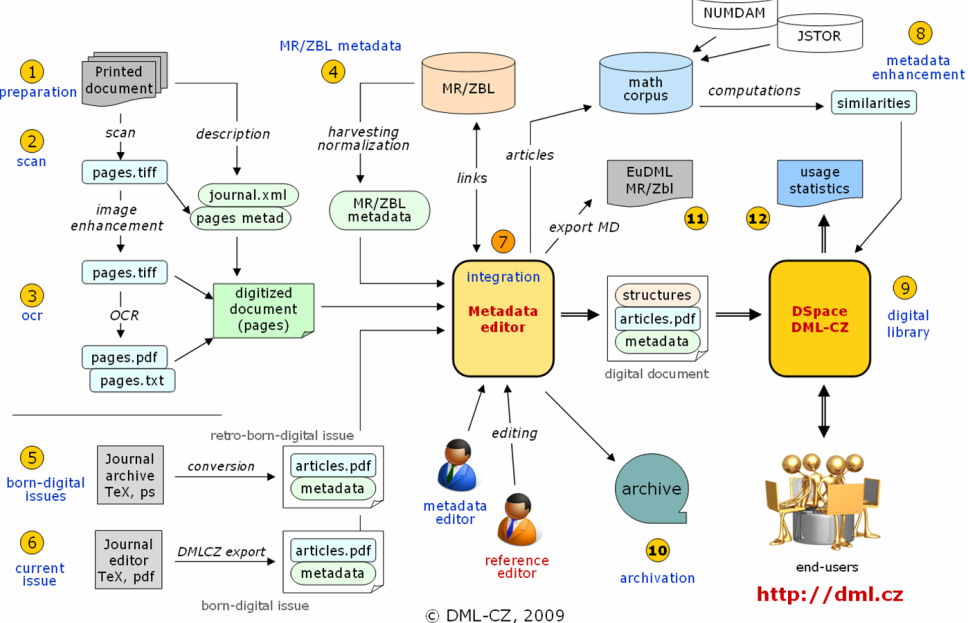
DML-CZ document engineering—data processing

DML-CZ challenges and lessons learned

DML-CZ, the Czech Digital Mathematics Library, now serves almost *300,000 pages of 30,000 math papers*. Challenges were

- *migration of existing workflows (retro-digital, retro-digital and born-digital) into the repository*
- negotiations with Google Scholar towards better visibility
- math indexing and search
- copyright and sustainability issues
- visualization
- space and processing demands
- . . . .

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

born-digital period: typesetting by TeX with export of [meta]data into digital library

world of authors: LaTeX, TeX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: $\rightarrow$ high automation to save costs

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

born-digital period: typesetting by TeX with export of [meta]data into digital library

world of authors: LaTeX, TeX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: → high automation to save costs

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

born-digital period: typesetting by TeX with export of [meta]data into digital library

world of authors: LaTeX, TeX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: → high automation to save costs

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations
(BookRestorer), OCR (FineReader, InftyReader),
two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad
formats, bitmap fonts of low resolution

born-digital period: typesetting by TeX with export of [meta]data
into digital library

world of authors: LaTeX, TeX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: → high automation to save costs

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

born-digital period: typesetting by TeX with export of [meta]data into digital library

world of authors: LaTeX, TeX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: $\rightarrow$ high automation to save costs

Document engineering 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

born-digital period: typesetting by TEX with export of [meta]data into digital library

world of authors: LATEX, TEX notation of mathematics

world of applications/data exchange: XML, MathML

big volumes: → high automation to save costs

Document engineering technologies and tools

6+ years of local (Brno, CZ) document engineering

Verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)

- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)

- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)

- gensim framework: similarity article computations aka document clustering (PhD research by Radim Řehůřek)

- web-based long distance metadata editing: web application metadata editor

Verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)

- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)

- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)

- gensim framework: similarity article computations aka document clustering (PhD research by Radim Řehůřek)

- web-based long distance metadata editing: web application metadata editor

Verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)

- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)

- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)

- gensim framework: similarity article computations aka document clustering (PhD research by Radim Řehůřek)

- web-based long distance metadata editing: web application metadata editor

Verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)

- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)

- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)

- gensim framework: similarity article computations aka document clustering (PhD research by Radim Řehůřek)

- web-based long distance metadata editing: web application metadata editor

Verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)

- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)

- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)

- gensim framework: similarity article computations aka document clustering (PhD research by Radim Řehůřek)

- web-based long distance metadata editing: web application metadata editor

DML-CZ workflow

① preparation

② scan

③ ocr

⑤ born-digital issues

⑥ current issue

Printed document

scan   *description*

pages.tiff

image enhancement

pages.tiff

OCR

pages.pdf

pages.txt

journal.xml

pages metad

digitized document (pages)

Journal archive TeX, ps   *conversion*

Journal editor TeX, pdf   *DMLCZ export*

retro-born-digital issue

articles.pdf

metadata

articles.pdf

metadata

born-digital issue

MR/ZBL metadata

④

MR/ZBL

harvesting normalization

MR/ZBL metadata

links

⑦ integration

Metadata editor

editing

metadata editor

reference editor

⑩ archivation

archive

NUMDAM

JSTOR

math corpus

computations

similarities

⑧ metadata enhancement

articles

EuDML MR/Zbl

export MD

⑪

⑫

structures

articles.pdf

metadata

digital document

usage statistics

DSpace DML-CZ

⑨ digital library

end-users

http://dml.cz

© DML-CZ, 2009

integration

Metadata editor

structures

articles.pdf

metadata

digital document

editing

metadata editor

reference editor

archive

10

archivation

issue

cles.pdf

etadata

cles.pdf

etadata

Metadata Editor http://editor.dml.cz

Web-based client-server tool allowing long-distance editing in any browser open source development (ICS MU) from scratch (Ruby) for [meta]data import, editing, validation, batch checking and correction.

To test, try
<http://editor.dml.cz:9129>,
admin/admin

Verified and proven technologies (in DML-CZ) (cont.)

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML

- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)

- Citation linking: CiteCrawl (BT Lukáš Lalinský)

- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)

- Visualization and browsing interface (MT Zuzana Nevěřilová)

Verified and proven technologies (in DML-CZ) (cont.)

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Citation linking: CiteCrawl (BT Lukáš Lalinský)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

Verified and proven technologies (in DML-CZ) (cont.)

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Citation linking: CiteCrawl (BT Lukáš Lalinský)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

Verified and proven technologies (in DML-CZ) (cont.)

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Citation linking: CiteCrawl (BT Lukáš Lalinský)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

Verified and proven technologies (in DML-CZ) (cont.)

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Citation linking: CiteCrawl (BT Lukáš Lalinský)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

Verified and proven technologies (in DML-CZ) (cont.)

Metadata (in RDF) visualisation, browsing: Visual Browser tool (MT Zuzana Nevěřilová) for [Eu]DML GUI.

Visual Browser visualisation

Verified and proven technologies (in DML-CZ) (cont.)

- batch digital signature of PDF: `pdfsign` (BT Peter Bočák)

- optimization of PDF: pdfopt (from `ghostscript` suite), `pdfsizeopt.py` (by Google sponsored Peter Szabó)

- PDF recompression using JBIG2: an application based on `jbig2enc`/Leptonica (started as BT by Radim Hatlapatka)

Verified and proven technologies (in DML-CZ) (cont.)

- batch digital signature of PDF: `pdfsign` (BT Peter Bočák)
- optimization of PDF: `pdfopt` (from `ghostscript` suite), `pdfsizeopt.py` (by Google sponsored Peter Szabó)
- PDF recompression using JBIG2: an application based on `jbig2enc`/Leptonica (started as BT by Radim Hatlapatka)

Verified and proven technologies (in DML-CZ) (cont.)

- batch digital signature of PDF: `pdfsign` (BT Peter Bočák)
- optimization of PDF: `pdfopt` (from `ghostscript` suite), `pdfsizeopt.py` (by Google sponsored Peter Szabó)
- PDF recompression using JBIG2: an application based on `jbig2enc/Leptonica` (started as BT by Radim Hatlapatka)

PDF tools
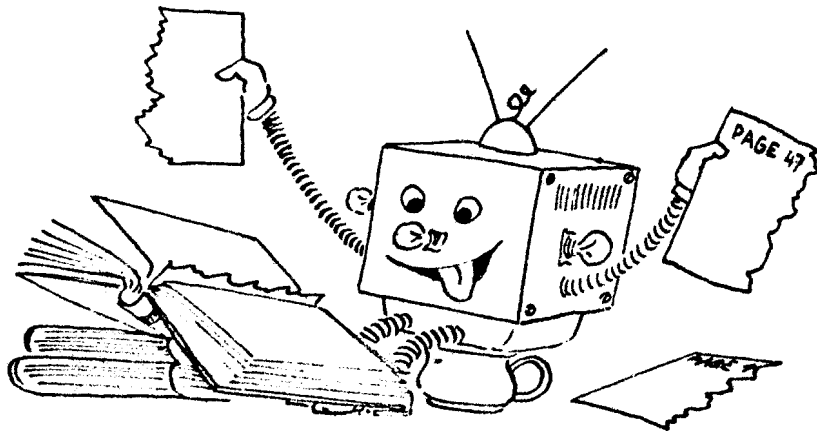
PDF recompressor

- Open-source tool written in Java for recompression of bitonal images
- Uses benefits of standard JBIG2 which is supported in PDF since version 1.4 (Acrobat 5)
- Uses improved `jbig2enc` with symbol coding used for text area
- Supports multi-page compression

PDF tools: PDF recompressor

```
┌─────────────────┐                 ┌─────────────────┐
│    Input PDF    │                 │    Output PDF   │
└─────────────────┘                 └─────────────────┘
         │                                   ▲
         ▼                                   │
┌─────────────────┐                 ┌─────────────────┐
│ Image extraction│                 │ Replacing images│
│                 │                 │     in PDF      │
└─────────────────┘                 └─────────────────┘
         │                                   ▲
         ▼                                   │
┌─────────────────┐       ┌─────────────────────────────┐
│ Jbig2enc encoder│──────▶│   Associating encoder       │
│                 │       │   output with image info    │
└─────────────────┘       └─────────────────────────────┘
```

PDF recompressor: input PDF

```
┌─────────────────┐          ┌─────────────────┐
│   Input PDF     │          │   Output PDF    │
└─────────────────┘          └─────────────────┘
         │                            ▲
         ▼                            │
┌─────────────────┐          ┌─────────────────┐
│ Image extraction│          │ Replacing images in PDF │
└─────────────────┘          └─────────────────┘
         │                            ▲
         ▼                            │
┌─────────────────┐          ┌─────────────────┐
│ Jbig2enc encoder│─────────▶│ Associating encoder     │
│                 │          │ output with image info  │
└─────────────────┘          └─────────────────┘
```
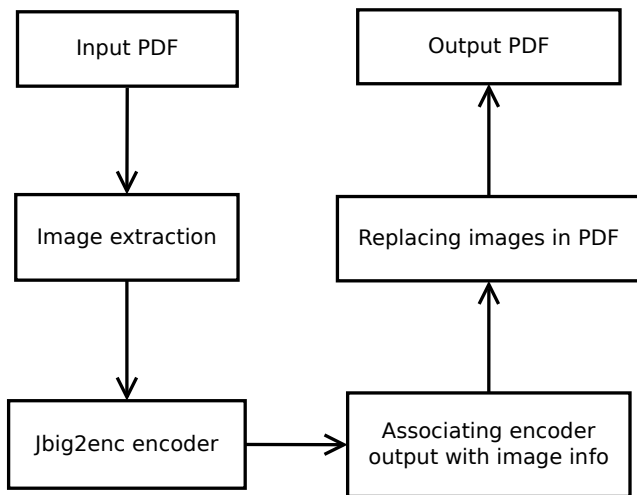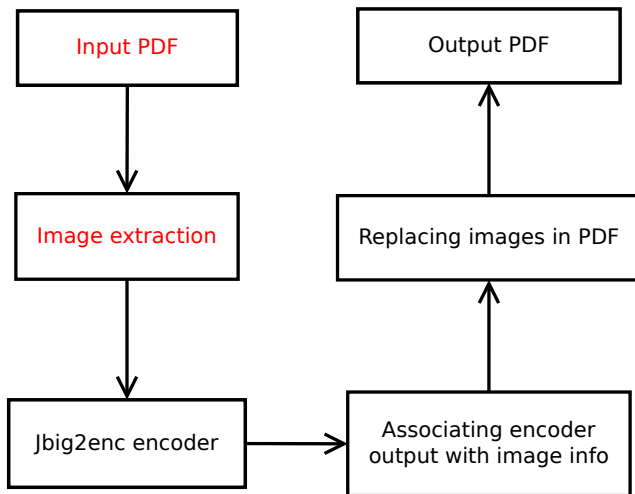
PDF recompressor: input PDF

```
27 0 obj << /Type/XObject
        /Subtype/Image
        /Name/im1
        /Length 47053
        /Width 2294
        /Height 3502
        /BitsPerComponent 1
        /ColorSpace/DeviceGray
        /Filter/CCITTFaxDecode
        /DecodeParms << /K -1
                /EndOfLine false
                /EncodedByteAlign false
                /Columns 2294
                /EndOfBlock true >>
        >>
        stream
        ...
        endstream
```

PDF recompressor via encoder `jbig2enc`

```
┌─────────────────┐              ┌─────────────────┐
│   Input PDF     │              │   Output PDF    │
└─────────────────┘              └─────────────────┘
         │                                ▲
         ▼                                │
┌─────────────────┐              ┌─────────────────────┐
│ Image extraction│              │Replacing images in PDF│
└─────────────────┘              └─────────────────────┘
         │                                ▲
         ▼                                │
┌─────────────────┐      ┌──────────────────────────┐
│ Jbig2enc encoder│─────▶│   Associating encoder     │
│                 │      │   output with image info  │
└─────────────────┘      └──────────────────────────┘
```
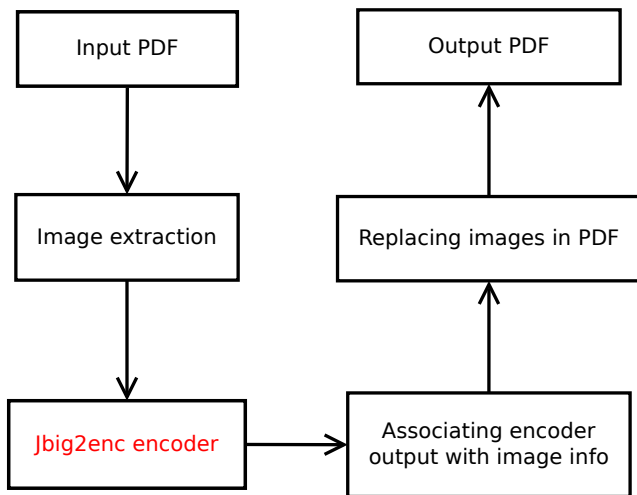
PDF recompressor: associating output with image info

```
┌─────────────────┐              ┌─────────────────┐
│    Input PDF    │              │   Output PDF    │
└─────────────────┘              └─────────────────┘
         │                                ▲
         ▼                                │
┌─────────────────┐              ┌─────────────────┐
│ Image extraction│              │Replacing images │
│                 │              │     in PDF      │
└─────────────────┘              └─────────────────┘
         │                                ▲
         ▼                                │
┌─────────────────┐   ┌──────────────────────────┐
│ Jbig2enc encoder│──▶│  Associating encoder     │
│                 │   │  output with image info  │
└─────────────────┘   └──────────────────────────┘
```
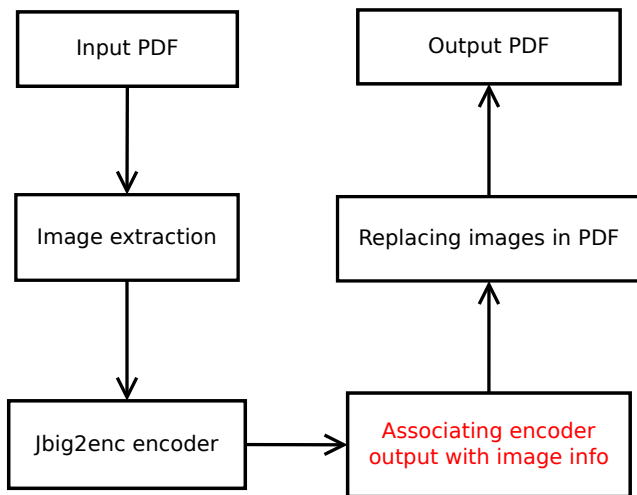
PDF recompressor: output PDF

```
┌─────────────────┐        ┌─────────────────┐
│   Input PDF     │        │   Output PDF    │
└────────┬────────┘        └────────▲────────┘
         │                          │
         ▼                          │
┌─────────────────┐        ┌─────────────────┐
│ Image extraction │        │ Replacing images in PDF │
└────────┬────────┘        └────────▲────────┘
         │                          │
         ▼                          │
┌─────────────────┐        ┌─────────────────────────┐
│ Jbig2enc encoder │──────▶│ Associating encoder     │
│                 │        │ output with image info  │
└─────────────────┘        └─────────────────────────┘
```

JBIG2 and `jbig2enc` basic principles

- Page segmented to several regions based on type of data (text, image, generic)

- For each region is used specific coding

- Text area segmented to connected components (symbols)

- For each new symbol is created a representive one and instances of this symbol are just pointers to this canonical representative

JBIG2 and `jbig2enc` basic principles

- Page segmented to several regions based on type of data (text, image, generic)
- For each region is used specific coding
- Text area segmented to connected components (symbols)
- For each new symbol is created a representive one and instances of this symbol are just pointers to this canonical representative

JBIG2 and `jbig2enc` basic principles

- Page segmented to several regions based on type of data (text, image, generic)

- For each region is used specific coding

- Text area segmented to connected components (symbols)

- For each new symbol is created a representive one and instances of this symbol are just pointers to this canonical representative

JBIG2 and `jbig2enc` basic principles

- Page segmented to several regions based on type of data (text, image, generic)
- For each region is used specific coding
- Text area segmented to connected components (symbols)
- For each new symbol is created a representive one and instances of this symbol are just pointers to this canonical representative

Improvement of `jbig2enc`—motivation

- Number of symbols recognized for a page is several times greater than of born digital documents

- Our improvement reduces size of output image in average for further 10 percent without visible loss

Improvement of `jbig2enc`—motivation

- Number of symbols recognized for a page is several times greater than of born digital documents
- Our improvement reduces size of output image in average for further 10 percent without visible loss

Improvement of `jbig2enc`

- Comparison of representative symbols
  - Two symbols are considered equivalent iff there is not found a big enough difference to form a line or a point

- Key idea: safe unification of two equivalent symbols to one

Improvement of `jbig2enc`

- Comparison of representative symbols
  - Two symbols are considered equivalent iff there is not found a big enough difference to form a line or a point
- Key idea: safe unification of two equivalent symbols to one

Improvement of `jbig2enc`

- Comparison of representative symbols
  - Two symbols are considered equivalent iff there is not found a big enough difference to form a line or a point
- Key idea: safe unification of two equivalent symbols to one

Image before and after compression

Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,

Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,

Image before and after compression: differences

Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,

Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,
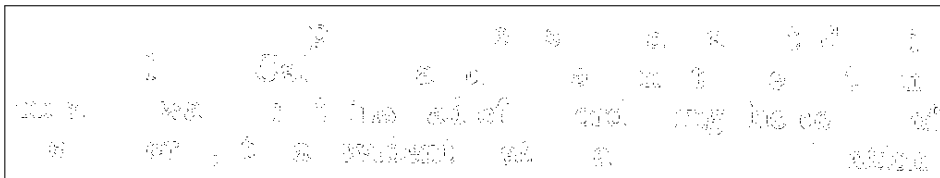
Image before and after compression: differences

PDF tools: `pdfsizeopt.py`

- Generic PDF optimizer written in Python by Péter Szabó (Google)

- Uses best practices and Unix tools to optimize size of PDF document (e.g. image compression, font unification)

- Uses `ghostscript`, `Multivalent`, `sam2p`, `pngout`, `jbig2enc`, ...

- Uses only generic coding of `jbig2enc`

- Images compressed using different compression methods and chooses one with the best result

PDF tools: `pdfsizeopt.py`

- Generic PDF optimizer written in Python by Péter Szabó (Google)
- Uses best practices and Unix tools to optimize size of PDF document (e.g. image compression, font unification)
- Uses `ghostscript`, `Multivalent`, `sam2p`, `pngout`, `jbig2enc`, . . .
- Uses only generic coding of `jbig2enc`
- Images compressed using different compression methods and chooses one with the best result

PDF tools: `pdfsizeopt.py`

- Generic PDF optimizer written in Python by Péter Szabó (Google)
- Uses best practices and Unix tools to optimize size of PDF document (e.g. image compression, font unification)
- Uses `ghostscript`, `Multivalent`, `sam2p`, `pngout`, `jbig2enc`,...
- Uses only generic coding of `jbig2enc`
- Images compressed using different compression methods and chooses one with the best result

PDF tools: `pdfsizeopt.py`

- Generic PDF optimizer written in Python by Péter Szabó (Google)
- Uses best practices and Unix tools to optimize size of PDF document (e.g. image compression, font unification)
- Uses `ghostscript`, `Multivalent`, `sam2p`, `pngout`, `jbig2enc`,...
- Uses only generic coding of `jbig2enc`
- Images compressed using different compression methods and chooses one with the best result

PDF tools: `pdfsizeopt.py`

- Generic PDF optimizer written in Python by Péter Szabó (Google)
- Uses best practices and Unix tools to optimize size of PDF document (e.g. image compression, font unification)
- Uses `ghostscript`, `Multivalent`, `sam2p`, `pngout`, `jbig2enc`,...
- Uses only generic coding of `jbig2enc`
- Images compressed using different compression methods and chooses one with the best result

Results: description of data used to create statistics

- PDF files of 11 journals retro-digitized in DML-CZ
- PDF files contain scanned text (bitonal page images originally compressed by CCITT-G4)
- Applied at PDF documents from digitized journal Archivum Mathematicum from years 1965–1991
- 6,641 pages in 665 papers in total

Results: description of data used to create statistics

- PDF files of 11 journals retro-digitized in DML-CZ
- PDF files contain scanned text (bitonal page images originally compressed by CCITT-G4)
- Applied at PDF documents from digitized journal Archivum Mathematicum from years 1965–1991
- 6,641 pages in 665 papers in total

Results: description of data used to create statistics

- PDF files of 11 journals retro-digitized in DML-CZ
- PDF files contain scanned text (bitonal page images originally compressed by CCITT-G4)

- Applied at PDF documents from digitized journal Archivum Mathematicum from years 1965–1991

- 6,641 pages in 665 papers in total

Results: description of data used to create statistics

- PDF files of 11 journals retro-digitized in DML-CZ
- PDF files contain scanned text (bitonal page images originally compressed by CCITT-G4)
- Applied at PDF documents from digitized journal `Archivum Mathematicum` from years 1965–1991
- 6,641 pages in 665 papers in total

Results: description of data used to create statistics

- PDF files of 11 journals retro-digitized in DML-CZ
- PDF files contain scanned text (bitonal page images originally compressed by CCITT-G4)
- Applied at PDF documents from digitized journal `Archivum Mathematicum` from years 1965–1991
- 6,641 pages in 665 papers in total

Results: different parts of PDFs

	Original PDF	After using PDF recompressor	After using `pdfsizeopt.py`	After using both
Total size (in kB)	7,123 (100%)	4,702 (66.01%)	3,962 (55.62%)	2,717 (38.14%)
Font data objects (in kB)	1,525 (100%)	1,525 (100%)	103 (6.74%)	103 (6.74%)
Image objects (in kB)	4,717 (100%)	1,915 (40.6%)	3,529 (74.83%)	1,904 (40.37%)
Other objects (in kB)	545 (100%)	926 (169.76%)	31 (5.63%)	411 (75.38%)

Results: single vs multi page PDF

Single page documents (655.83 MB in total)			
	By using PDF recompressor	By using `pdfsizeopt.py`	By using both
Saved globally	77.37%	52.22%	46.68% (396 MB)
Saved in image and other objects	70.46%	60.30%	52.97%

Multi page documents (723.47 MB in total)			
	By using PDF recompressor	By using `pdfsizeopt.py`	By using both
Saved globally	66.01%	55.62%	38.14% (276 MB)
Saved in image and other objects	53.99%	67.66%	44.00%

Summary

- Verified complex DML-CZ digitization workflow and proven technologies and tools for math DL

- PDF size reduction of sixtytwo percent of original already CCITT-G4 compressed PDFs using PDF recompressor with improved `jbig2enc` and `pdfsizeopt.py`

- EuDML: Towards wordwide digital mathematical library, based on DML-CZ know-how and tools developed at Masaryk University during last ≈ 6 years

- DML workshop series, join us at DML 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

Summary

- Verified complex DML-CZ digitization workflow and proven technologies and tools for math DL
- PDF size reduction of sixtytwo percent of original already CCITT-G4 compressed PDFs using PDF recompressor with improved `jbig2enc` and `pdfsizeopt.py`
- EuDML: Towards wordwide digital mathematical library, based on DML-CZ know-how and tools developed at Masaryk University during last $\approx$ 6 years
- DML workshop series, join us at DML 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011
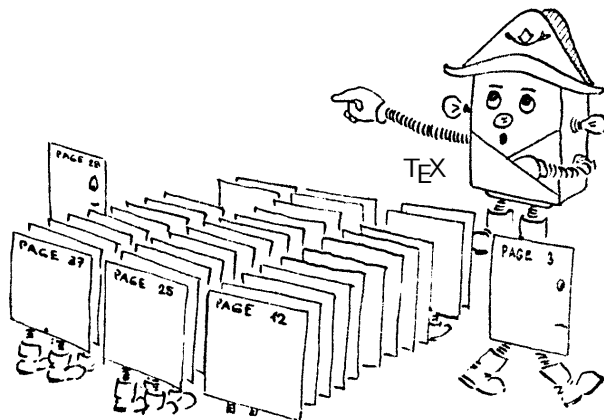
Summary

- Verified complex DML-CZ digitization workflow and proven technologies and tools for math DL
- PDF size reduction of sixtytwo percent of original already CCITT-G4 compressed PDFs using PDF recompressor with improved `jbig2enc` and `pdfsizeopt.py`
- EuDML: Towards wordwide digital mathematical library, based on DML-CZ know-how and tools developed at Masaryk University during last ≈ 6 years
- DML workshop series, join us at DML 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

Summary

- Verified complex DML-CZ digitization workflow and proven technologies and tools for math DL
- PDF size reduction of sixtytwo percent of original already CCITT-G4 compressed PDFs using PDF recompressor with improved `jbig2enc` and `pdfsizeopt.py`
- EuDML: Towards wordwide digital mathematical library, based on DML-CZ know-how and tools developed at Masaryk University during last ≈ 6 years
- DML workshop series, join us at DML 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

Yes, you can!

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further

- Optimize subimage lookup and storage in PDF recompressor

- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity

- Design alternative and novel user interfaces for the digital library

- Improve metadata validation procedures in ME

- Interfaces for export and conversion for projects on European or worldwide levels

- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense

- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

Future work

- Adding OCR tools to PDF recompressor to increase compression ratio of bitonal images even further
- Optimize subimage lookup and storage in PDF recompressor
- Pursue research in mathematical document classification, math indexing and retrieval, OCR for math, document similarity
- Design alternative and novel user interfaces for the digital library
- Improve metadata validation procedures in ME
- Interfaces for export and conversion for projects on European or worldwide levels
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense
- Cooperation "wanted!" for problems above, `fixfont`, math OCR

End of the talk
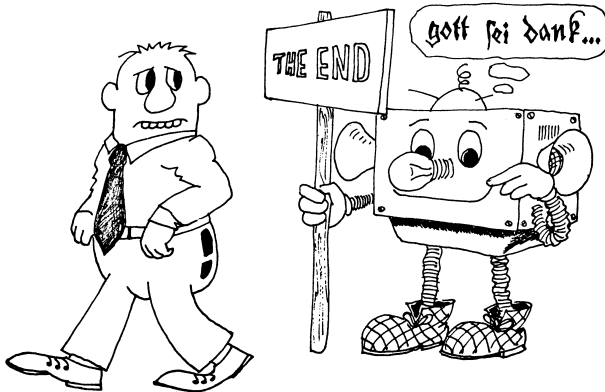
Questions? Comments? Cooperation offers?

References

Patrice Y. Simard, Henrique S. Malvar, James Rinker, Erin Renshaw:
A Foreground/Background Separation Algorithm for Image Compression.

Dan Bloomberg.
Leptonica [online, cit. 2010-11-04].
<http://www.leptonica.com/>.

L. Bottou and P. Haffner and P. G. Howard and P. Simard and Y. Bengio and Y. Le Cun:
High Quality Document Image Compression with DjVu
<http://leon.bottou.org/papers/bottou-98>.

R. Hatlapatka:
Website of the PDF recompression project.
<http://nlp.fi.muni.cz/projekty/eudml/pdfRecompression/>.

Adam Langley:
Jbig2enc [online, cit. 2010-11-04].
<http://github.com/agl/jbig2enc/>.

Péter Szabó:
Optimizing PDF output size of TEX documents [online, cit. 2010-11-04].
<http://code.google.com/p/pdfsizeopt/>.

DML-CZ team.
Materials about DML-CZ, project publications [online, cit. 2010-11-04].
<http://project.dml.cz/documents.html>.

References (cont.)

EuDML team.
EuDML project info [online, cit. 2010-11-04].
`<http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503>`.

EuDML team.
EuDML webpage [online, cit. 2010-11-04].
`<http://eudml.eu/>`.

EuDML at MU team.
EuDML at MU project info [online, cit. 2010-11-04].
`<http://nlp.fi.muni.cz/projekty/eudml/>` or `<http://www.muni.cz/research/projects/10067>`.