Document Engineering for a Digital Library

PDF recompression using JBIG2 and other optimization of PDF

Petr Sojka and Radim Hatlapatka

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<sojka@fi.muni.cz>, <208155@fi.muni.cz>

DocEng 2010, Manchester, UK, September 22nd

Outline and two take-off messages

1. Motivation, vision of PubMed Central for Mathematics
2. Complexity of digitization workflow of The Czech Digital Mathematics Library DML-CZ
3. Document engineering technologies and tools for DML-CZ and EuDML
4. Tools developed (PDF Re-compressor et al.)
5. Results: already compressed 2-layer bitonal PDF squeezed to 38%
6. Summary, conclusions and future work

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.

Several attempts to fund development of WDML on world-wide (NSF/de Moore foundation) and European level (FP5, FP6) were not successful. Finally three year Pilot B project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010. The strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY 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.

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.

Several attempts to fund development of WDML on world-wide (NSF/de Moore foundation) and European level (FP5, FP6) were not successful. Finally three year Pilot B project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010. The

strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY 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.

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.

Several attempts to fund development of WDML on world-wide (NSF/de Moore foundation) and European level (FP5, FP6) were not successful. Finally three year Pilot B project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010. The strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY 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.

Decade of the vision of WDML as PubMed 4 Math

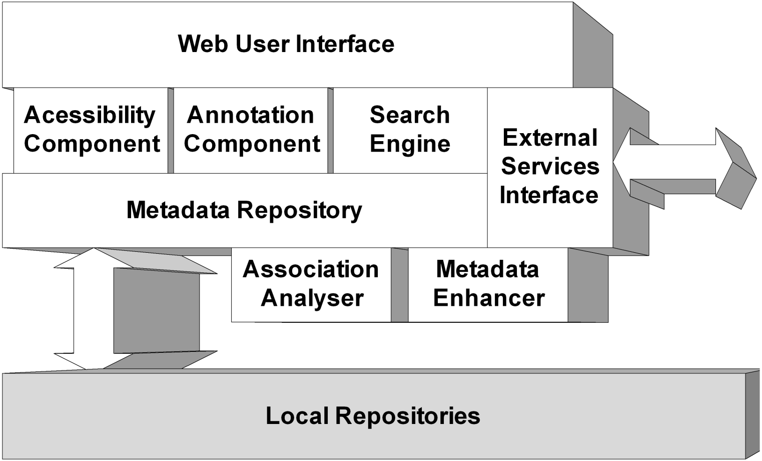
In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Several attempts to fund development of WDML on world-wide (NSF/de Moore foundation) and European level (FP5, FP6) were not successful. Finally three year Pilot B project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010. The strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY 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.

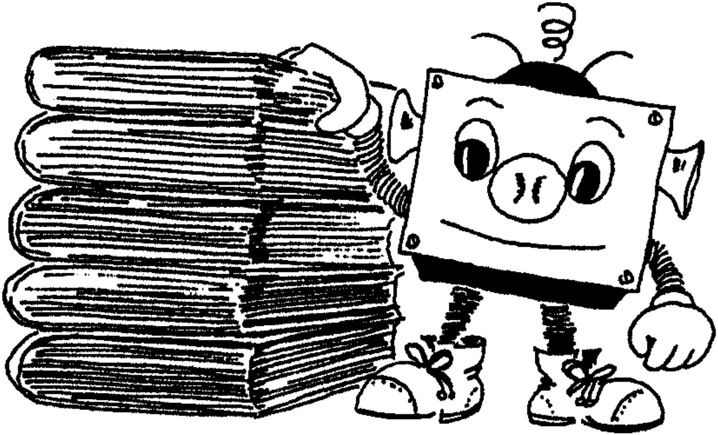
Example of DML-CZ: up and running digital mathematic library <http://dml.cz> with nearly 30,000 papers (300,000 pages). For more, see (who, what, browse, browse similar, how to search).

Bottom up—from building bricks of regional repositories

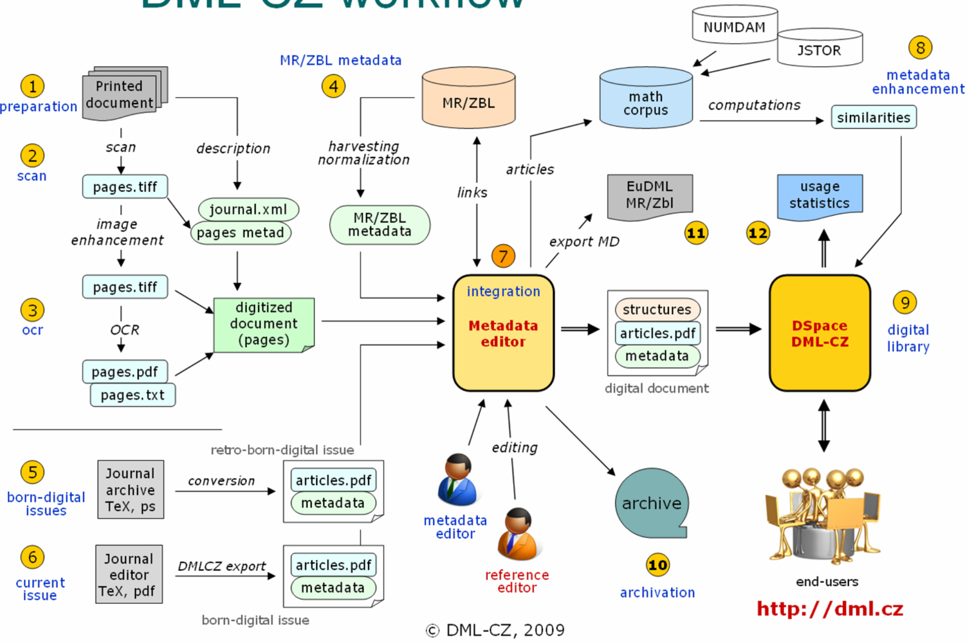
As DML content providers serve mostly publisher's or regional DML repositories as The Czech Digital Mathematics Library DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,. . . : aggregating content from local repositories to build the bigger (global?) DML.

Example of DML-CZ: up and running digital mathematic 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

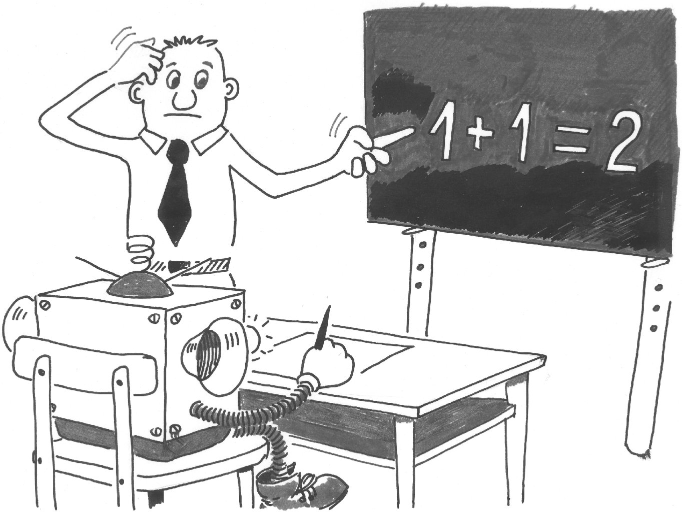


DML-CZ workflow

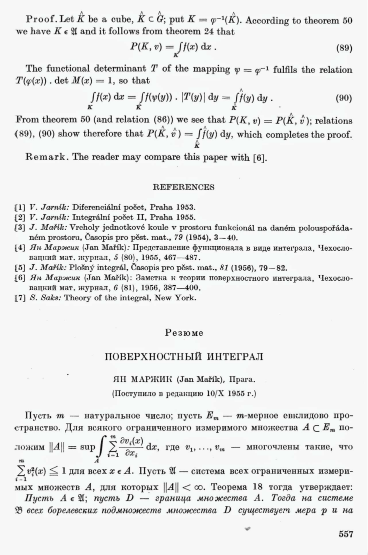


Take care! “God is in the details.” (Mies van der Rohe)

Challenges of Math handling: OCR, indexing, search. . .



DML-CZ—data: scientific math published in CZ/SK



Proof. Let $\hat{K}$ be a cube, $\hat{K} \subset \hat{G}$; put $K = \varphi^{-1}(\hat{K})$. According to theorem 50 we have $K \in \mathfrak{A}$ and it follows from theorem 24 that

$$P(K, v) = \int_K f(x)\,\mathrm{d}x\,. \tag{89}$$

The functional determinant T of the mapping $\psi = \varphi^{-1}$ fulfils the relation $T(\varphi(x))\,.\,\det M(x) = 1$, so that

$$\int_K f(x)\,\mathrm{d}x = \int_{\hat{K}} f(\psi(y))\,.\,|T(y)|\,\mathrm{d}y = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y\,. \tag{90}$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y$, which completes the proof.

Remark. The reader may compare this paper with [6].

REFERENCES

[1] *V. Jarník:* Diferenciální počet, Praha 1953.
[2] *V. Jarník:* Integrální počet II, Praha 1955.
[3] *J. Mařík:* Vrcholy jednotkové koule v prostoru funkcionál na daném polouspořádaném prostoru, Časopis pro pěst. mat., *79* (1954), 3—40.
[4] *Ян Маржик* (Jan Mařík): Представление функционала в виде интеграла, Чехословацкий мат. журнал, *5* (80), 1955, 467—487.
[5] *J. Mařík:* Plošný integrál, Časopis pro pěst. mat., *81* (1956), 79—82.
[6] *Ян Маржик* (Jan Mařík): Заметка к теории поверхностного интеграла, Чехословацкий мат. журнал, *6* (81), 1956, 387—400.
[7] *S. Saks:* Theory of the integral, New York.

Резюме

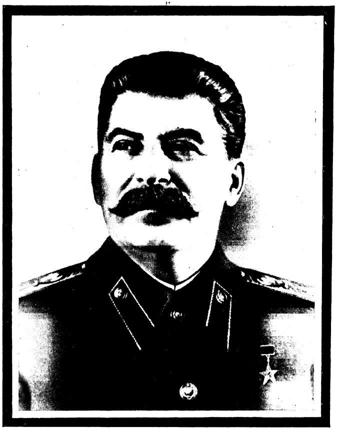
ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЖИК (Jan Mařík), Прага.

(Поступило в редакцию 10/X 1955 г.)

Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^m \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^m v_i^2(x) \leqq 1$ для всех $x \in A$. Пусть $\mathfrak{A}$ — система всех ограниченных измеримых множеств A, для которых $\|A\| < \infty$. Теорема 18 тогда утверждает: *Пусть $A \in \mathfrak{A}$; пусть D — граница множества A. Тогда на системе $\mathfrak{B}$ всех борелевских подмножеств множества D существует мера p и на*

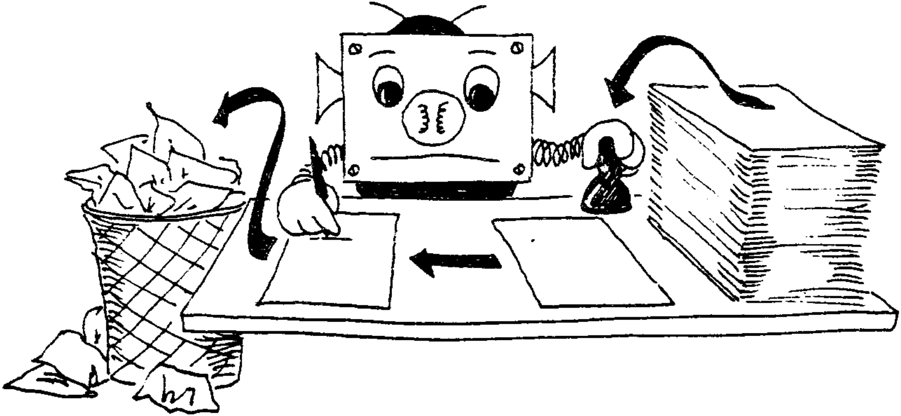
557



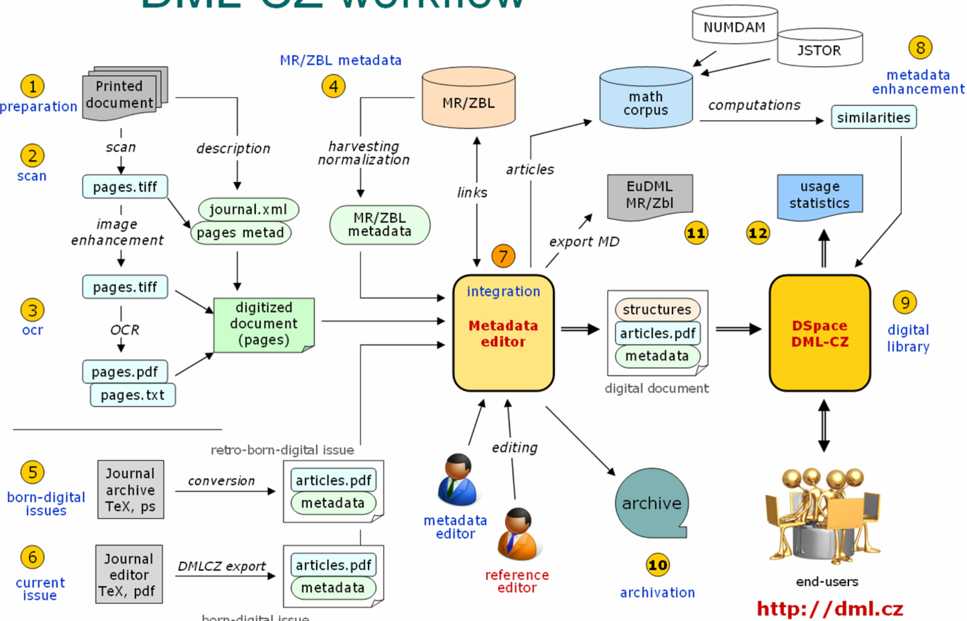
ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

Document engineering—from paper to digital *workflow*

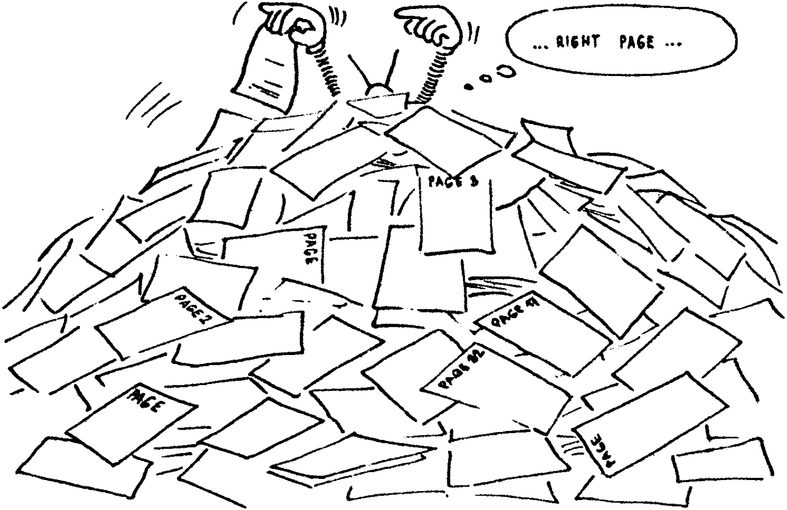


DML-CZ workflow



© DML-CZ, 2009

DML-CZ document engineering—data processing



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: → 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

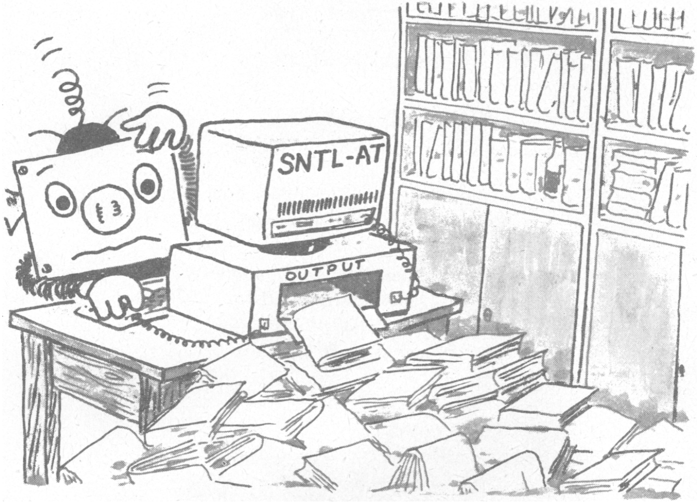
born-digital period: typesetting by TEX with export of [meta]data into digital library

world of authors: LATEX, TEX notation of mathematics

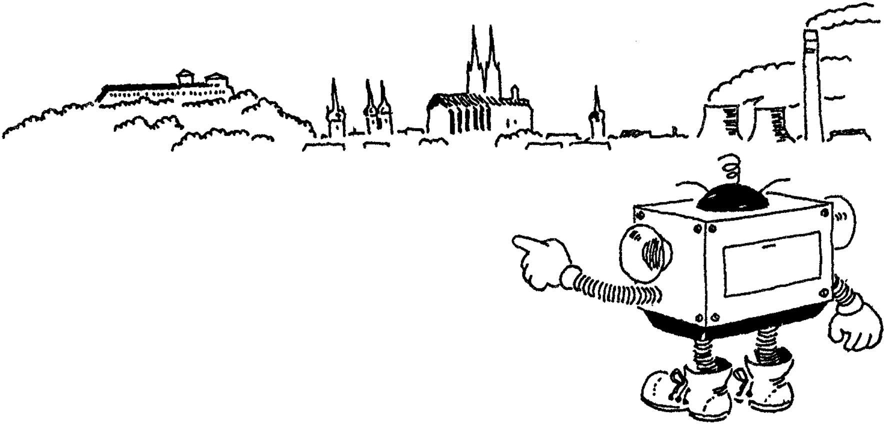
world of applications/data exchange: XML, MathML

big volumes: → high automation to save costs

Document engineering 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 (Radim Řehůřek)
- gensim framework: similarity article computations (machine learning research with Radim Řehůřek)

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 (Radim Řehůřek)
- gensim framework: similarity article computations (machine learning research with Radim Řehůřek)

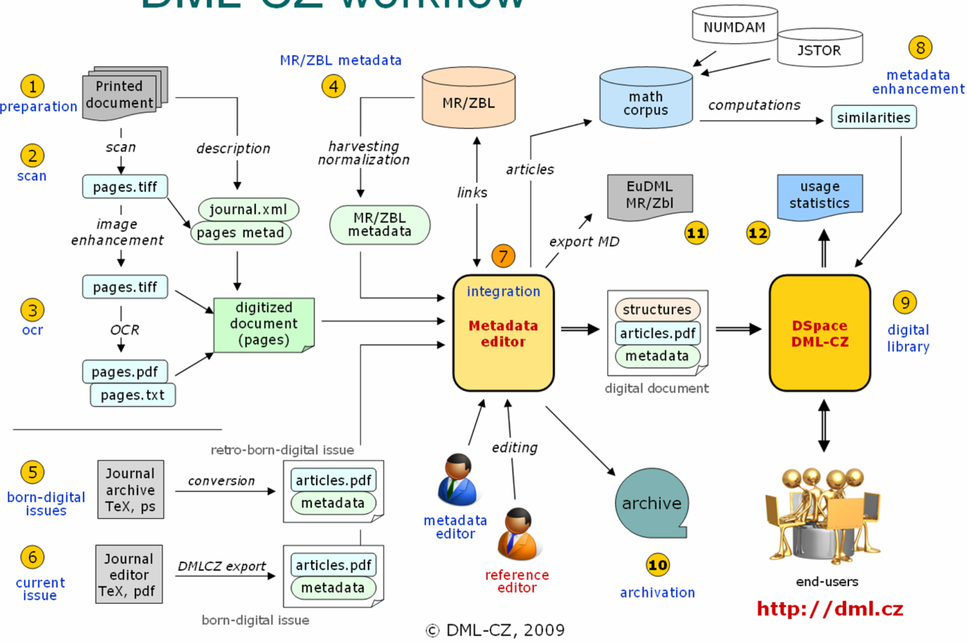
Verified and proven technologies (in DML-CZ)

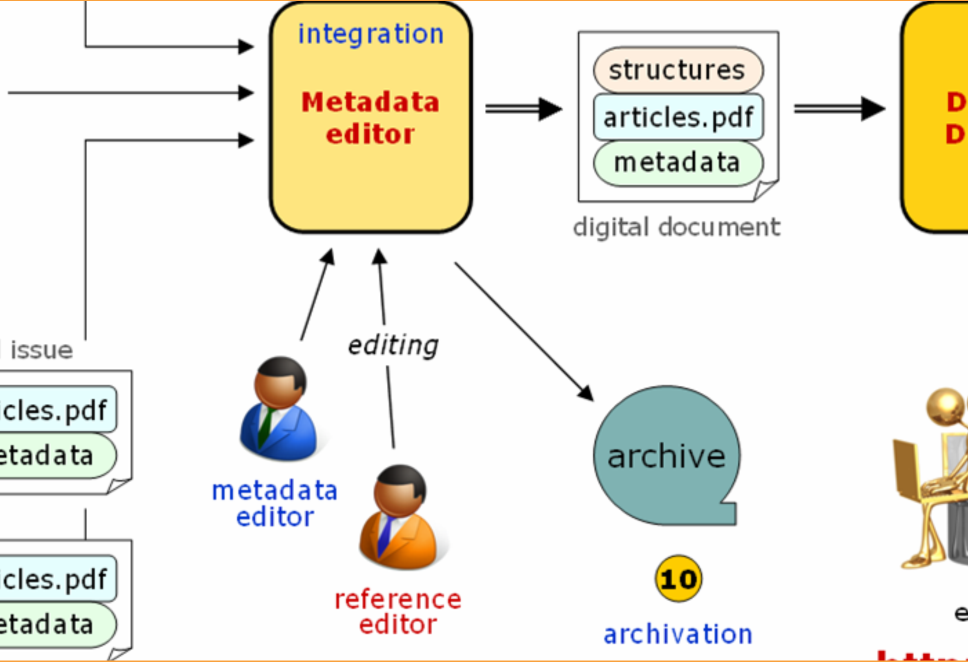
- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT Vystrčil)
- Pre-MSC era papers' automated classification by MSC (Radim Řehůřek)
- gensim framework: similarity article computations (machine learning research with Radim Řehůřek)

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 (Radim Řehůřek)
- gensim framework: similarity article computations (machine learning research with Radim Řehůřek)

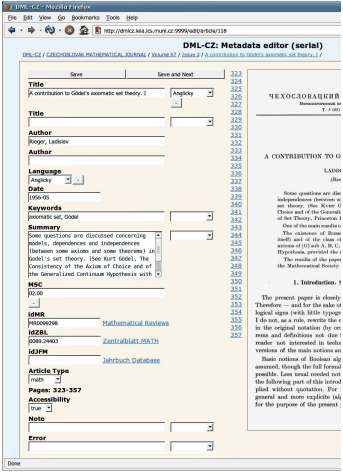
DML-CZ workflow




integration
Metadata editor
structures
articles.pdf
metadata
digital document
D
D
issue
cles.pdf
etadata
cles.pdf
etadata
editing
metadata editor
reference editor
archive
10
archivation
e

Metadata Editor http://editor.dml.cz

Web-based client-server tool
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



DML-CZ: Metadata editor (serial)

DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 07 / Issue 3 / A contribution to Gödel's axiomatic set theory. I /

Title: A contribution to Gödel's axiomatic set theory. I — Anglicky

Title:

Author: Rieger, Ladislav

Author:

Language: Anglicky

Date: 1956-05

Keywords: axiomatic set, Gödel

Summary: Some questions are discussed concerning models, dependences and independences (between some axioms and some theorems) in Godel's set theory. (See Kurt Gödel, The Consistency of the Axiom of Choice and of the Generalized Continuum Hypothesis with

MSC: 02.00

idMR: MR0099298

idZBL: 0089.24403

idJFM:

Article Type: math

Pages: 323-357

Accessibility: true

Note:

Error:

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)

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)

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

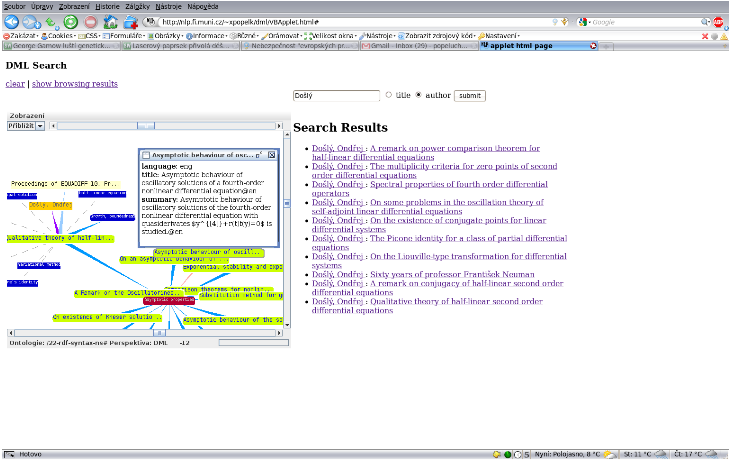
- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Citation linking: CiteCrawl (BT Lukáš Lalinský)
- Born-digital publishing system [for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

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)

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

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


Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda
http://nlp.fi.muni.cz/~xpopelk/dml/VBApplet.html#
Google
Zakázat Cookies CSS Formuláře Obrázky Informace Různé Orámovat Velikost okna Nástroje Zobrazit zdrojový kód Nastavení
George Gamow luští genetick...
Laserový paprsek přivolá déš...
Nebezpečnost "evropských pr...
Gmail - Inbox (29) - popeluch...
applet html page
Zobrazení
Přiblížit
Proceedings of EQUADIFF 10, Pr...
ipal solution
half-linear equation
Došlý, Ondřej
Growth, boundedness
Qualitative theory of half-lin...
variational method
ne's identity
Asymptotic behaviour of oscill...
On an asymptotic behaviour of ...
Exponential stability and expo
Comparison theorems for nonlin...
A Remark on the Oscillatorines...
Substitution method for g
Asymptotic properties
On existence of Kneser solutio...
Asymptotic behaviour of the so
Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -12
Hotovo
Nyní: Polojasno, 8 °C
St: 11 °C
Čt: 17 °C

DML Search

clear | show browsing results

Došlý ○ title ● author submit

Asymptotic behaviour of osc...

language: eng
title: Asymptotic behaviour of oscillatory solutions of a fourth-order nonlinear differential equation@en
summary: Asymptotic behaviour of oscillatory solutions of the fourth-order nonlinear differential equation with quasiderivates $y^{[4]}+r(t)f(y)=0$ is studied.@en

Search Results

- Došlý, Ondřej : A remark on power comparison theorem for half-linear differential equations
- Došlý, Ondřej : The multiplicity criteria for zero points of second order differential equations
- Došlý, Ondřej : Spectral properties of fourth order differential operators
- Došlý, Ondřej : On some problems in the oscillation theory of self-adjoint linear differential equations
- Došlý, Ondřej : On the existence of conjugate points for linear differential systems
- Došlý, Ondřej : The Picone identity for a class of partial differential equations
- Došlý, Ondřej : On the Liouville-type transformation for differential systems
- Došlý, Ondřej : Sixty years of professor František Neuman
- Došlý, Ondřej : A remark on conjugacy of half-linear second order differential equations
- Došlý, Ondřej : Qualitative theory of half-linear second order differential equations

Verified and proven technologies (cont.): PDF

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

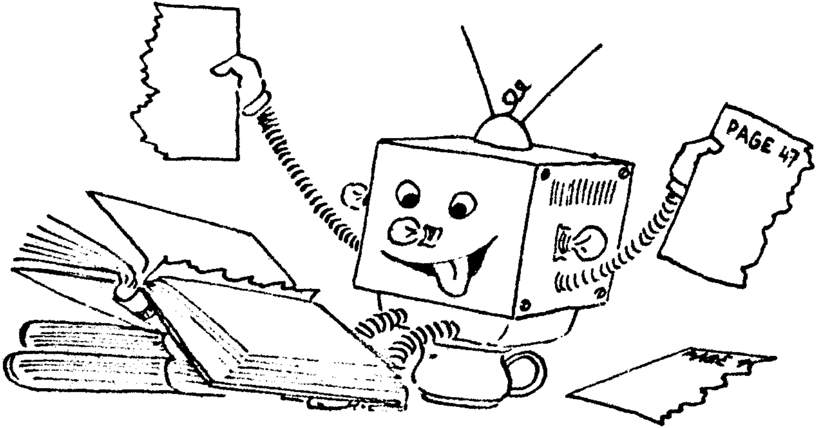
Verified and proven technologies (cont.): PDF

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

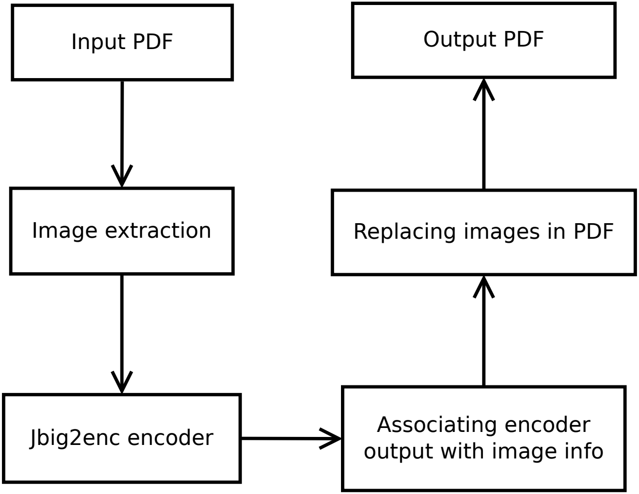
Verified and proven technologies (cont.): PDF

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

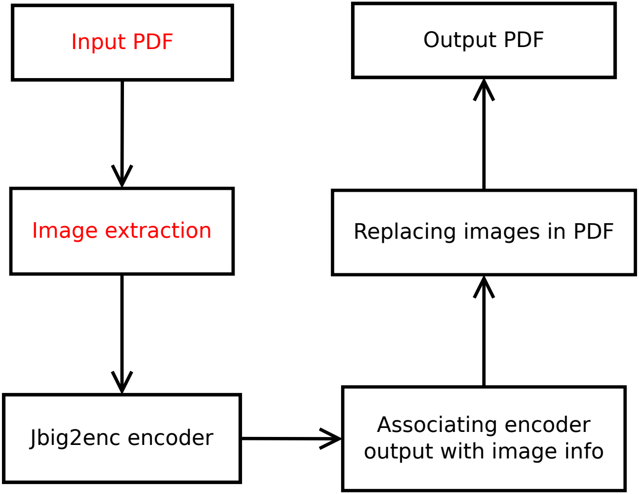
PDF tools



PDF tools: PDF re-compressor



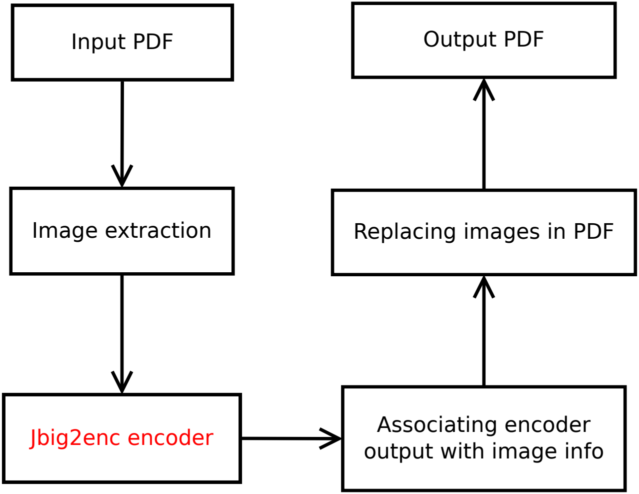
PDF re-compressor: input PDF



PDF re-compressor: 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 re-compressor via encoder jbig2enc



Jbig2enc and Leptonica

- Open-source JBIG2 encoder developed by Adam Langley, commissioned by Google [Books]
- Open-source library Leptonica, developed by Dan Bloomberg, is used for manipulation with images and bitmaps of symbols
- Symbols (bitmaps of connected pixels) are encoding using a chosen bitmap as representant for each symbol and putting pointers to this representant
- Supports output in format suitable for PDF

Jbig2enc and Leptonica

- Open-source JBIG2 encoder developed by Adam Langley, commissioned by Google [Books]
- Open-source library Leptonica, developed by Dan Bloomberg, is used for manipulation with images and bitmaps of symbols
- Symbols (bitmaps of connected pixels) are encoding using a chosen bitmap as representant for each symbol and putting pointers to this representant
- Supports output in format suitable for PDF

Jbig2enc and Leptonica

- Open-source JBIG2 encoder developed by Adam Langley, commissioned by Google [Books]
- Open-source library Leptonica, developed by Dan Bloomberg, is used for manipulation with images and bitmaps of symbols
- Symbols (bitmaps of connected pixels) are encoding using a chosen bitmap as representant for each symbol and putting pointers to this representant
- Supports output in format suitable for PDF

Jbig2enc and Leptonica

- Open-source JBIG2 encoder developed by Adam Langley, commissioned by Google [Books]
- Open-source library Leptonica, developed by Dan Bloomberg, is used for manipulation with images and bitmaps of symbols
- Symbols (bitmaps of connected pixels) are encoding using a chosen bitmap as representant for each symbol and putting pointers to this representant
- Supports output in format suitable for PDF

Modification of jbig2enc

- Compare all templates (representative symbols) with the same size for finding equivalence on symbols
    - two templates are considered equivalent if there is not found big enough accumulation of differences
    - we look for accumulations in shapes such as points or lines
- Unify equivalent symbols

Modification of jbig2enc

- Compare all templates (representative symbols) with the same size for finding equivalence on symbols
    - two templates are considered equivalent if there is not found big enough accumulation of differences
    - we look for accumulations in shapes such as points or lines
- Unify equivalent symbols

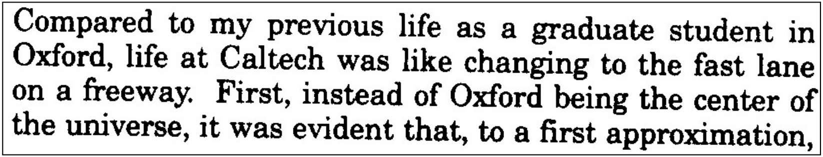
Modification of jbig2enc

- Compare all templates (representative symbols) with the same size for finding equivalence on symbols
  - two templates are considered equivalent if there is not found big enough accumulation of differences
  - we look for accumulations in shapes such as points or lines
- Unify equivalent symbols

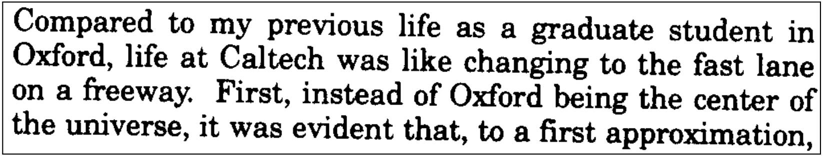
Modification of jbig2enc

- Compare all templates (representative symbols) with the same size for finding equivalence on symbols
    - two templates are considered equivalent if there is not found big enough accumulation of differences
    - we look for accumulations in shapes such as points or lines
- Unify equivalent symbols

Image before and after compression

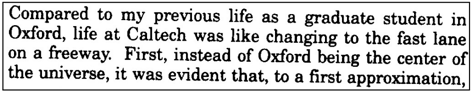


Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,

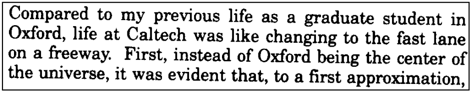


Compared to my previous life as a graduate student in Oxford, life at Caltech was like changing to the fast lane on a freeway. First, instead of Oxford being the center of the universe, it was evident that, to a first approximation,

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,

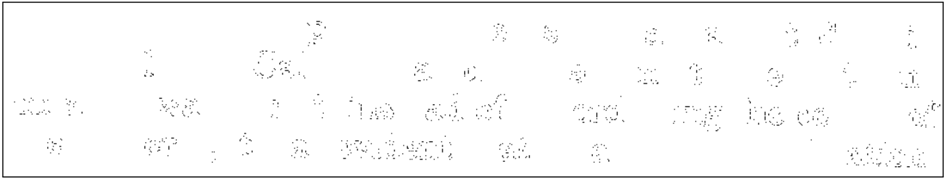
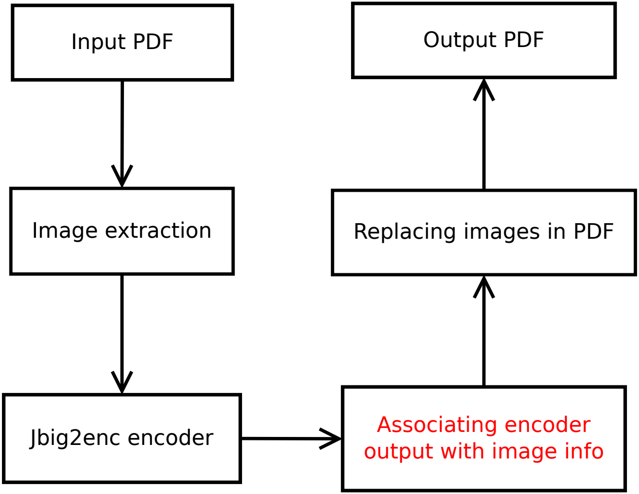
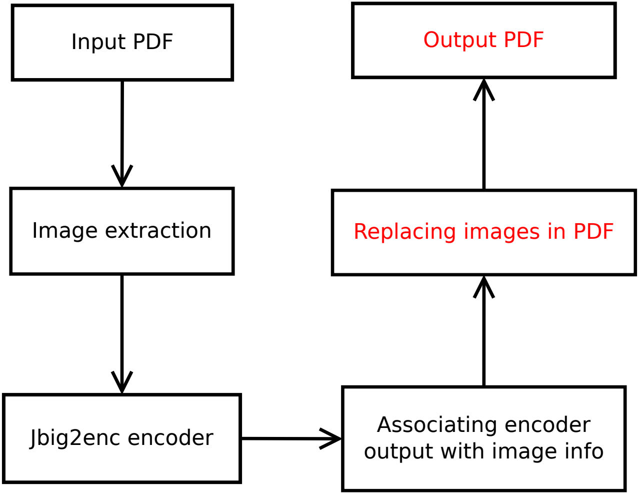


Image before and after compression: differences

PDF re-compressor: associating output with image info



PDF re-compressor: output PDF



PDF re-compressor: PDF image encoded using JBIG2

```
2 0 obj << /DecodeParms
           << /JBIG2Globals 1 0 R >>
           /Width 2294
           /BitsPerComponent 1
           /Height 3502
           /Filter /JBIG2Decode
           /Subtype /Image
           /Length 34336
           /ColorSpace /DeviceGray
           /Type /XObject
        >>
        stream
        ...
        endstream
```

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: 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

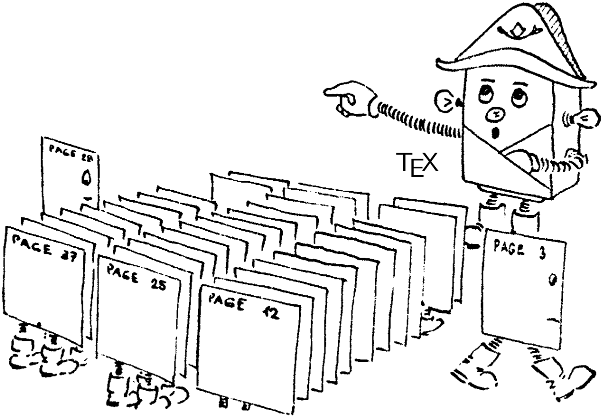
Summary

- 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

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

Yes, you can!



Future work

- Adding OCR tools to PDF re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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 re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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 re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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 re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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.

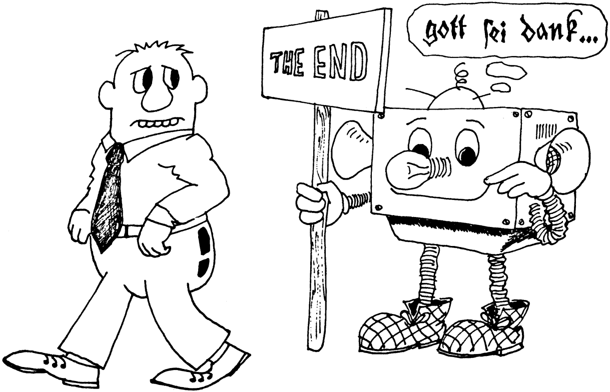
Future work

- Adding OCR tools to PDF re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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 re-compressor to increase compression ratio of bitonal images even further.
- Optimize subimage lookup and storage in PDF re-compressor.
- 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.

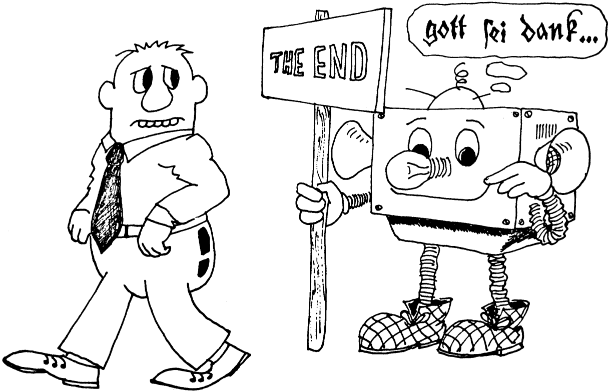
End of the talk



Questions? Comments?

▸ Continue by pictorial summary if time permits.

End of the talk



Questions? Comments?

▸ Continue by pictorial summary if time permits.

References

Dan Bloomberg.
Leptonica [online, cit. 2010-09-09].
<http://www.leptonica.com/>.

Adam Langley:
Jbig2enc [online, cit. 2010-09-09].
<http://github.com/agl/jbig2enc/>.

Péter Szabó:
Optimizing PDF output size of TEX documents [online, cit. 2010-09-09].
<http://code.google.com/p/pdfsizeopt/>.

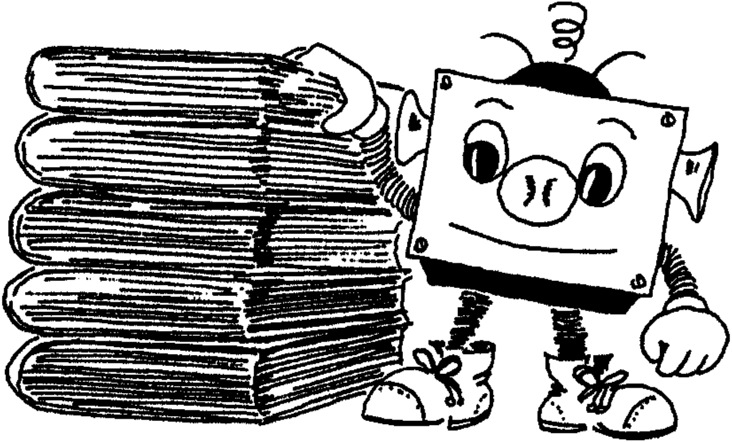
DML-CZ team.
Materials about DML-CZ, project publications [online, cit. 2010-09-09].
<http://project.dml.cz/documents.html>.

EuDML team.
EuDML project info [online, cit. 2010-09-09].
<http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503>.

EuDML team.
EuDML webpage [online, cit. 2010-09-09].
<http://eudml.eu/>.

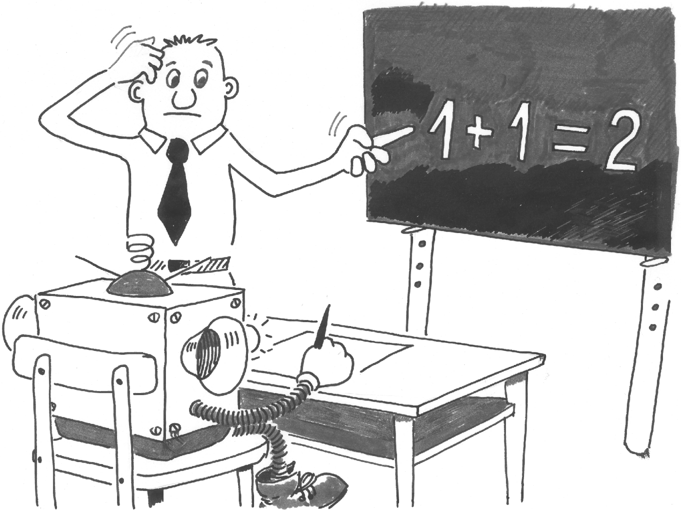
EuDML at MU team.
EuDML at MU project info [online, cit. 2010-09-09].
<http://nlp.fi.muni.cz/projekty/eudml/> or <http://www.muni.cz/research/projects/10067>.

From paper to digital processing

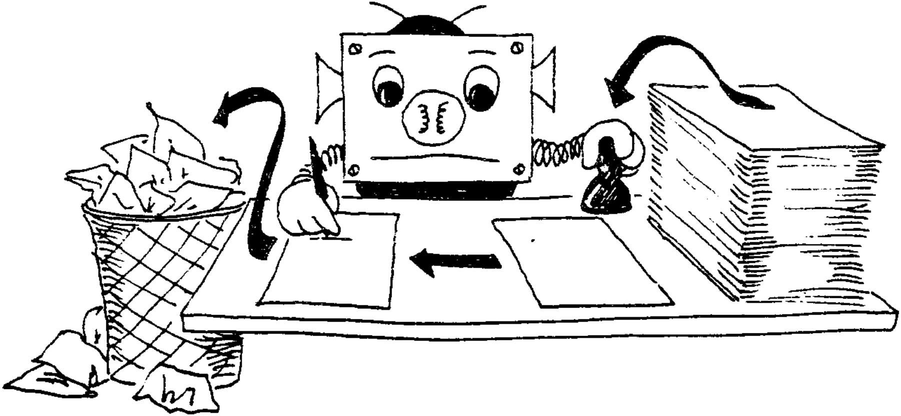


Information overload in globalized scientific world

Information overload also in specific domains (mathematics)

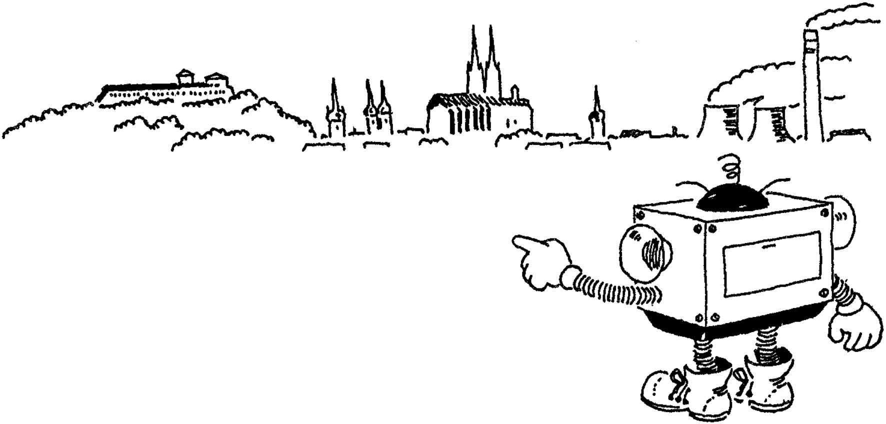


Document Engineering (DocEng): from paper to digital workflow

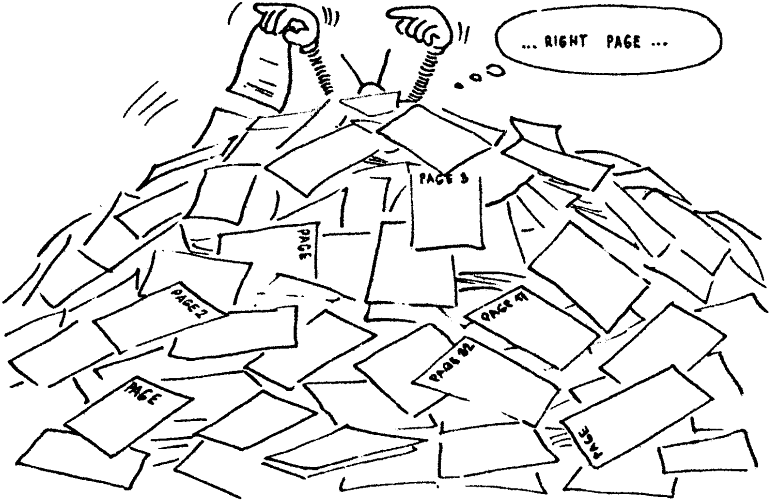


DocEng: retro-digitization, digital library development

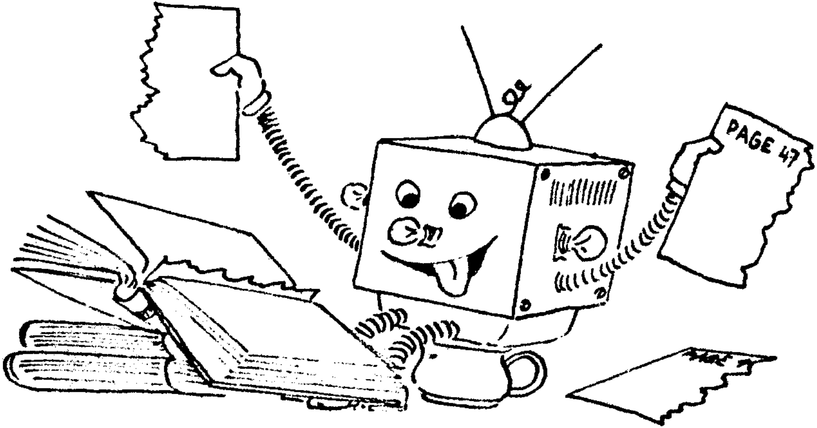
DocEng for specific/local (Brno, CZ) purposes



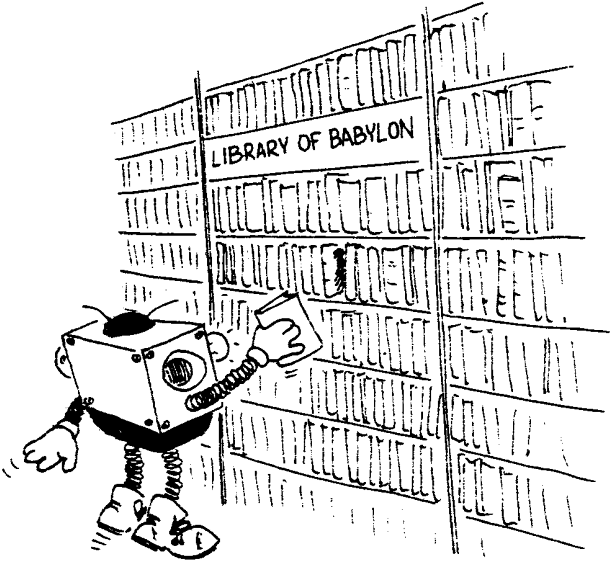
DocEng in DML-CZ: new workflows and data processing



DocEng in DML-CZ: new tools

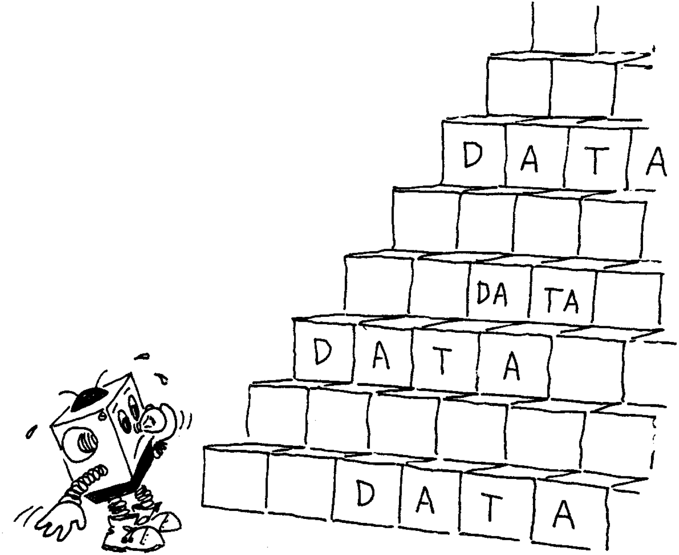


'Bottom up' deployment towards EU or worldwide scale

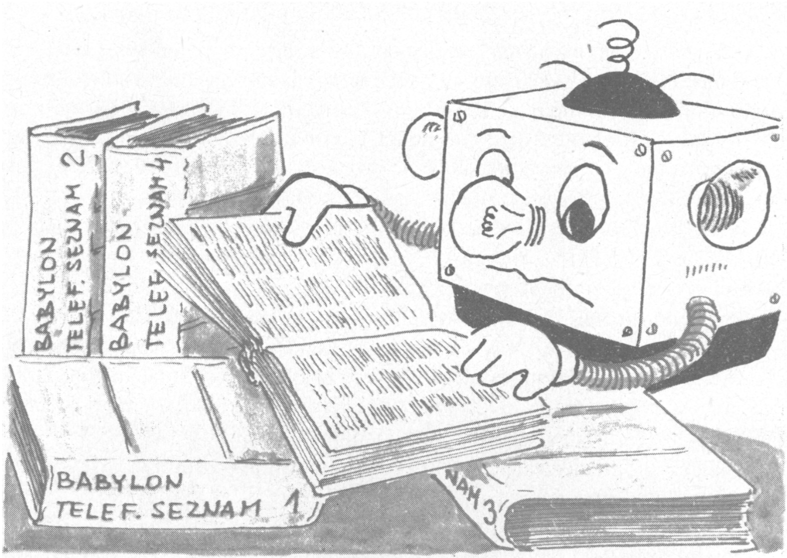


The European Digital Mathematics Library: EuDML

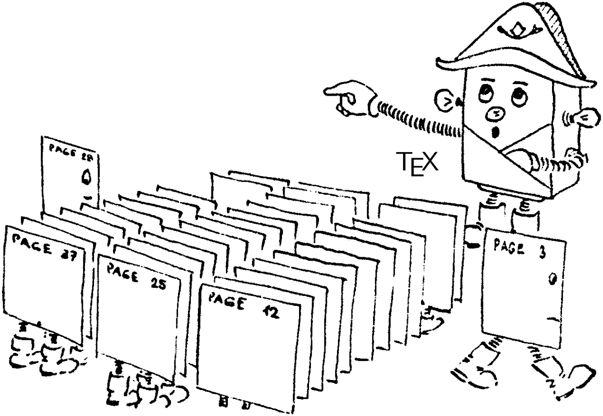
EuDML: from local data collections to the virtual DL



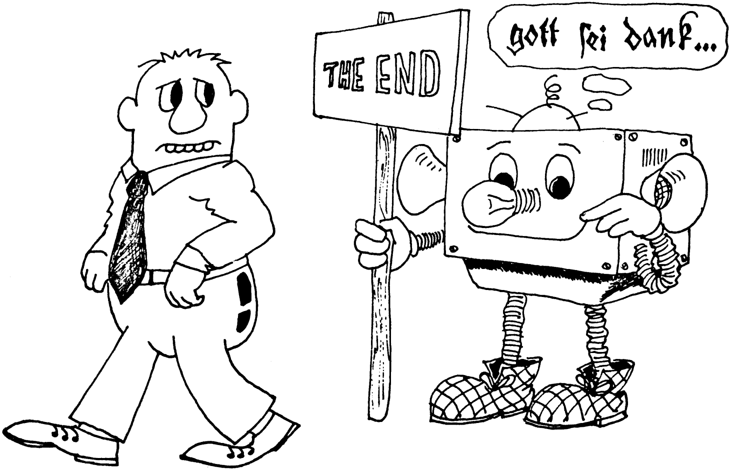
DocEng for EuDML: scalable tools development



Yes, you can! You can have visibility, scalability, similarity fulltext metrics, 38% of original size PDFs,. . .



End of talk overview



DjVu

- What is DjVu? DjVu is open document format (alternative to PDF) designed to store scanned text especially with text, line drawings and photographs.
- How are images compressed? Image is divided into three images (foreground, background and mask).
- Background and foreground images are compressed using a wavelet-based compression algorithm named IW44.
- What is JB2? It is compression method similar to JBIG2 used for compression of mask image.

DjVu

- What is DjVu? DjVu is open document format (alternative to PDF) designed to store scanned text especially with text, line drawings and photographs.
- How are images compressed? Image is divided into three images (foreground, background and mask).
- Background and foreground images are compressed using a wavelet-based compression algorithm named IW44.
- What is JB2? It is compression method similar to JBIG2 used for compression of mask image.

DjVu

- What is DjVu? DjVu is open document format (alternative to PDF) designed to store scanned text especially with text, line drawings and photographs.
- How are images compressed? Image is divided into three images (foreground, background and mask).
- Background and foreground images are compressed using a wavelet-based compression algorithm named IW44.
- What is JB2? It is compression method similar to JBIG2 used for compression of mask image.

DjVu

- What is DjVu? DjVu is open document format (alternative to PDF) designed to store scanned text especially with text, line drawings and photographs.
- How are images compressed? Image is divided into three images (foreground, background and mask).
- Background and foreground images are compressed using a wavelet-based compression algorithm named IW44.
- What is JB2? It is compression method similar to JBIG2 used for compression of mask image.

DjVu

- What is DjVu? DjVu is open document format (alternative to PDF) designed to store scanned text especially with text, line drawings and photographs.
- How are images compressed? Image is divided into three images (foreground, background and mask).
- Background and foreground images are compressed using a wavelet-based compression algorithm named IW44.
- What is JB2? It is compression method similar to JBIG2 used for compression of mask image.

DjVu and JB2 – How is image segmented?

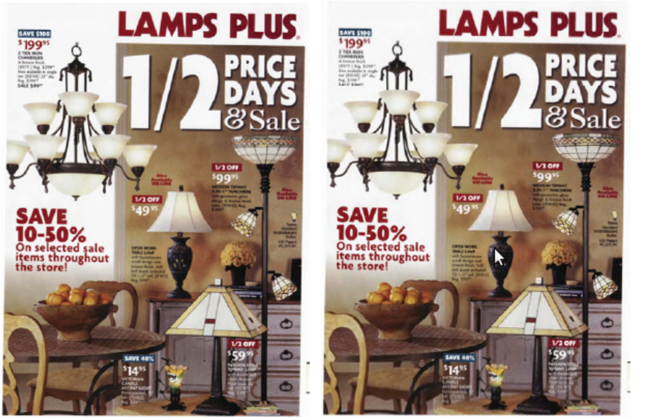


Figure: Image before (on the left) and after compression (on the right) [?]

DjVu and JB2 – How is image segmented? (cont.)

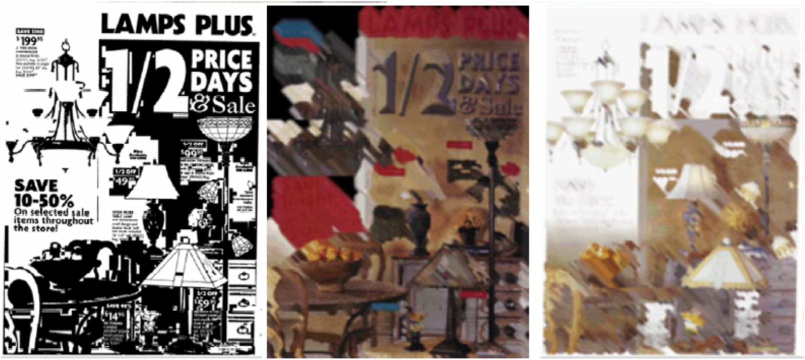


Figure: DjVu image components of the image shown at previous slide; left to right: Mask, Foreground and Background [?]