Digitisation Workflow in the Czech Digital Mathematics Library

PETR SOJKA

Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

Abstract

Experience in setting up a workflow from scanned images of mathematical writings into a fully fledged mathematical library is described on the example of the project Czech Digital Mathematics Library DML-CZ. An overview of the whole process is given, with detailed description of production steps involving scanned image processing and optical character recognition. Experience gained, lessons learned and tools prepared during development of DML-CZ are described. DML-CZ now serves over 25,600 articles (275,000 digitised pages) to the public.

Keywords: digital mathematical library, mathematical knowledge representation, digitisation workflow, optical character recognition, OCR, retro-digitisation, DML-CZ

Viva la Workflows! (Carole Goble [6])

1 Motivation

Digital Library business has moved from data/files centered processing towards process-oriented *workflows*. Workflows enact the machinery of building and running a digital library. Instead of running simple tools and mirroring file repositories more subtle solutions have to be devised: data curatorship changes to workflow curatorship and *services*.

There are communities and systems that start to dominate in some thematic areas: PubMed Central (PMC) is one such system in the medical domain, speeding up research and author's citation indexes in the area. Unfortunately, only domains where global initial funding was available took advantages of the platforms established and tools and workflows developed. In the PMC case, journal publishers are now eager to join the club, and authors enjoy global topical ontology-based search. Researchers send their papers only to journals available in PMC as this leverages their citation indexes. However, the realization of the dream of a World Digital Mathematics Library [7] is yet to come.

We report on the experience gained, lessons learned and tools prepared during the development of a digitisation workflow for The Czech Digital Mathematics Library DML-CZ project. The aim of the project approved for the five years period 2005–2009 was to digitize the relevant mathematical literature published in the Czech lands. It comprises periodicals, selected monographs and conference proceedings from the nineteenth century up until currently produced mathematical publications. It has been launched and is readily available on dml.cz, serving more than 25,600 articles on 275,000 pages to the public.

The general workflow of the project, shown on Figure 1 on the following page, reflects different types of acquired input data:

full digitisation from print work starts from a paper copy;

full digitisation from bitmap image work starts from an electronic bitmap of pages;

retro-born-digital work starts from an electronic version of the document (usually in POSTSCRIPT or PDF);

born-digital workflow of the journal production is enriched with an automated export of data for the digital library.

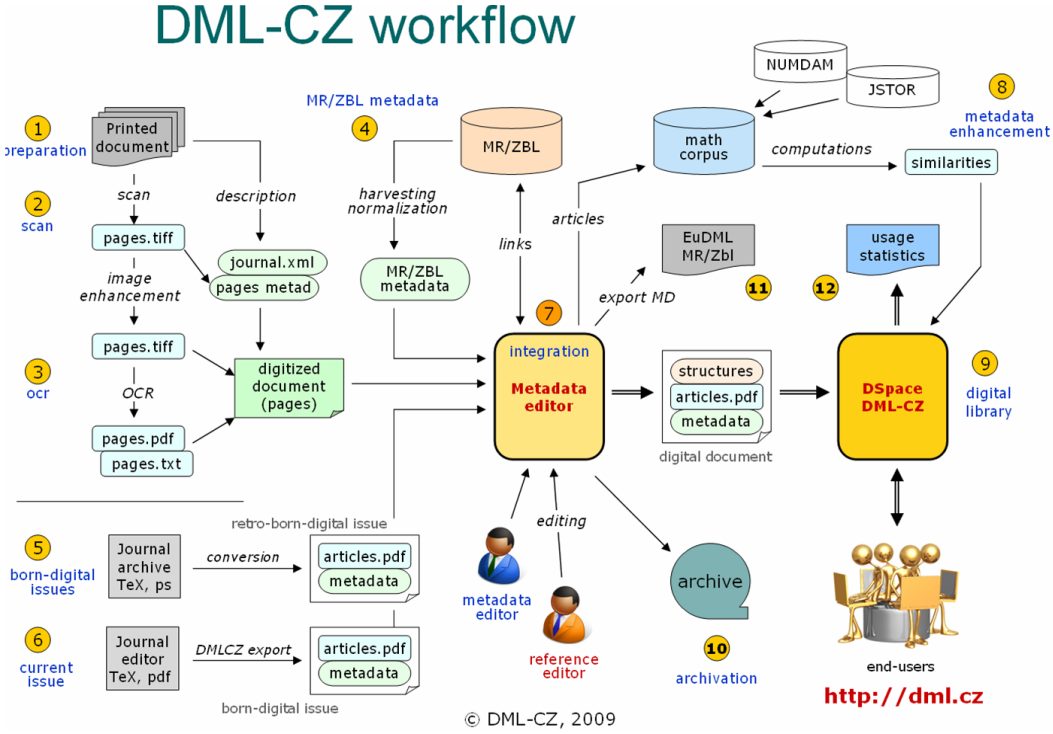

Figure 1: DML-CZ top-level workflow scheme

Within the project, several general purpose tools have been developed:

1. scripting of transformation pipes of scanned images,
2. DML-CZ OCR workflow allowing recognition of scanned mathematical documents,
3. web-based Metadata Editor [1],
4. tools for classification of mathematical documents and measuring their similarity [14];
5. workflow for born-digital publication production with direct export of metadata for DML [15] and
6. plenty of other smaller tools like: extensions to LUCENE engine allowing indexing of mathematics, batch PDF stamper for digital signing of produced PDF, an optimizer recompressing image objects in PDF with the new JBIG compression filter supported by Adobe since PDF specification version 1.6 (Adobe Reader 5) or batch article PDF generation with titlepage by XeLaTeX.

In the following sections we describe part (steps 2, 3, 7 and 8 in Figure 1) of our digitisation workflow with the hope that they can be used by similar projects or even in other domains.

We are all apprentices in a craft where no-one ever becomes a master. (Ernest Hemingway)

2 Scanning and Image Transformations

Processing of scanned images is aimed at final delivery of 600 DPI bi-tonal images suitable for quality OCR and a fine print. This is the quality recommended by the Committee on Electronic Information and Communication (CEIC) and used for example by JSTOR and NUMDAM. Images from the Göttingen Digitisation Centre (GDZ) and images scanned in the Digitisation Centre of the Library of Academy of Sciences, Czech Republic prior to the project DML-CZ have bi-tonal 400 DPI quality. The difference is visible, and leads to a higher OCR error rate. We strongly support the recommendation to scan with a resolution of at least 600 DPI.

We perform our new scans at 600 DPI with 4-bit depth, 'having a depth/space' for geometrical and other transformations done on images before binarization. Scanning is carried out in Digitisation Centre of the Library of Academy of Sciences in Jenštejn near Prague on the Zeutschel OS 7000 A2 book scanners.

The primary scans in TIFF format are archived for a possible future reprocessing if needed. We use BOOK RESTORER™ image restoration software by i2S for interactive and batch image processing in an uncompressed TIFF format. Operations performed on images are:

1. *geometrical correction* as narrowing the baselines and widths of the same characters on the same line;
2. *cropping* the page to cut out the speckles at page borders;
3. *blur filter*, 3×3 pixels, to eliminate one or two pixel size variations;
4. *binarisation* with manually adjusted parameters for every batch (usually journal volume);
5. *despeckle filter*, with both white and black spotting, 3×3 pixels;
6. *publish/export*: processed TIFFs are stored being compressed by the Lempel-Ziv-Welsh method for compressing grayscale and the G4 one for binarized images to speed up further processing (OCR) and to save space.

Both the order of these steps and the parameter adjustments for images of different quality are very important. For the data from GDZ, slightly different operations are needed as the input files are already bi-tonal and some filters are applicable only on grayscale images.

Step 1 employs the algorithms that allow perspective correction of a scanned image. As most of the material to digitize cannot be cut, we scan 2-up page spreads, making the text size non-uniform even when trying to flatten the spread by pane of glass. Book Restorer can also flatten the lighting across the scanned spead. For more details of this step see [2, page 2].

Step 2 crops the unnecessary border parts of the page shot.

Step 3 aims at better binarisation and despeckling by unsharping the shapes in the image.

Step 4 is necessary as most OCR engines work on bi-tonal images. It may be left to the high-quality OCR engine—clever thresholding starts to be a standard part of OCR programs [18], or perform it ourselves adaptively based on OCR feedback [13].

Step 5 is inserted to remove small impurities in the image.

Step 6 is the final step: image is stored as LZW-compressed grayscale or G4-compressed bi-tonal TIFF.

For the lower resolution data from GDZ, slightly different operations are needed as the input files are already bi-tonal (e.g. we did upscaling before unsharping) and because some filters are applicable only on grayscale images.

It is wise to differentiate processing of pages with grayscale images (e.g. photos) so that they are not degraded by image filters suitable for text. To avoid possible difficulties in the later steps it is important from the very beginning to carefully check image quality before proceeding with the remaining steps. At least automated procedures that check the technical metadata (e.g. `tiffinfo`) and image quality (pixel width and height) has to be the part of quality assurance. Metrics of compressibility by the JBIG2 encoder were used to trigger quality checks.

There is a tradeoff between price of image cleanup and quality results within constraints of digitisation budget. When acquiring a craftmanship in good image editing software, results very close to (or even better than) the original could be achieved [17]. These hand made touches are usually beyond the budget of most digitisation projects, where the highest degree of automation is needed to reduce the cost of digitisation. In DML-CZ, we have prepared [12] set of batches of typical transformation procedures to be used by BOOK RESTORER™ operators to achieve the best price/effort ratio.

Fine-tuning of operations on the pixel level pays back in the following step: the OCR.

The road to wisdom?
Well, it's plain and simple to express:
Err and err and err again,
but less and less and less. (Piet Hein)

3 Optical Character Recognition – DML-CZ OCR

To have papers indexed we need to get full text from page bitmaps by the process of optical character recognition. Also, we need to recognize logical page numbers located in every TIFF, to link the page images to article metadata.

Tests with various OCR programmes showed that no single one gives acceptable results for mathematical content, with character error rates often above 10% (counting wrong character positions and font types as errors too). For text recognition, FINEREADER by ABBYY® gave the best results, whereas for the structural recognition of mathematics InftyReader [24] had impressive results.

The FINEREADER software development kit (SDK for Windows version 8.1) was used to develop a part of the system for the location and recognition of page numbers, and a batch system DML-CZ OCR [19, 22] which takes sequences of TIFF images and produces two-layered one page PDFs (with invisible full-texts behind the images). The processing starts with the recognition of languages used in every paragraph, and then blocks are recognized again with a special setting (language dictionaries used) for every given block of text. With such a fine-tuning of parameters, we are able to achieve a one percent character error rate [22].

Among solutions and software evaluated on plain texts, FINEREADER gave the best results, but it has no support for the recognition of mathematical expressions. Texts without recognized maths may be sufficient for basic indexing and search. However, it is not surprising that omitting maths matters when the full texts are used for such tasks as automated text classification and categorization or for computing paper similarity [23]. Therefore we strive to enhance the state-of-the-art possibilities for mathematical OCR.

Neither ABBYY® nor Google responded positively on the near future of math OCR development plans—mathematics is only a small market niche for them. On the other hand, developers of the INFTYREADER system [24] were willing to gradually improve their support for European languages, MATHML and LaTeX export filters and to enrich their recognized database of mathematical symbols.

We found that setting the parameters of the OCR engine (language, word-list consultation) influences the precision significantly. We trained FINEREADER on the type cases used at the printer where journals were typeset.

At the end of extensive experiments, we developed a method of OCR processing consisting of several phases, both in FINEREADER and Infty. Processing using FINEREADER consist of the following:

1. A page or block of text is recognised for the first time using a universal setup (non-language specific). A histogram of character bigrams and trigrams from words with lengths greater than three is created.

2. The computed histogram of the text block is compared [5] to the histograms created from the journal data during the training phase for all languages used (English, French, Russian, German and Czech). Perl module `Lingua::Ident` is used. Block with bibliography is detected by different algorithms and is treated differently.

3. Page or block of text is processed for the second time with parameters optimised for recognised 'language' in previous step and saved as a two-layer PDF (with text layer used for searching, indexing and similarity computation).

Recognition of mathematical formulae in FINEREADER is not satisfactory, however. The only suitable tool for this domain that we have found and experimented with is INFTY. INFTY's new PDF import capability is very significant to us: it will allow to import our current FINEREADER's two-layer PDFs, use the text part only, throw away badly recognized maths and to detect and recognize maths expressions. A new INFTY version that combines FINEREADER's technology (OCR voting [9]) is in preparation. In the meantime,

1. PDF is passed to INFTYREADER and results are stored in the INFTY Markup Language (IML) and in LaTeX (Human readable LaTeX).

2. IML is postprocessed by a home-grown programme in JAVA to fix recognition errors of some of the accented characters that INFTY does not yet have in its glyph database.

Using the process outlined above we have managed to decrease the character error rate from an initial 11.35% (universal language setup of FineReader) to an average 0.98% character error rate. [10, 11, 20] The whole processing is fully automated after initial font recognition and language detection training. The error rate may be further decreased when INFTY's character database is semiautomatically enriched when processing a new journal.

When in doubt, use brute force. (Ken Thompson)

4 Text Postprocessing and Metadata Enhancements

The OCR step is followed by further text processing, and its results are used for editing of metadata and references.

4.1 Metadata Editor

The Metadata Editor (ME) [1, 4] has gradually developed into a fully-fledged and efficient web application, `https://editor.dml.cz`, that allows simultaneous remote editing according to assigned structured access rights. It supports two levels of actions. On the first one the operator editing the data is provided with page thumbnails so that he can visually check the completeness, scan the quality and configuration of the articles, easily shuffle the pages and cut or merge articles if necessary. On the other level the operator can check the automatically imported metadata, edit and complete them. An integral part of the ME is the module for administration of authority files with authors' names. It enables the most suitable version of the name for the DML-CZ to be selected and to match it with all its other versions.

We consider bibliographical references as important metadata of every paper. Their availability makes it possible to use professional systems like CrossRef® for cross-publisher citation linking. The work starts from OCR of the text, in which a block of references is found. Citations are tagged by a script based on regular expressions written for the citation style of every journal. The operator then checks, edits and approves the list of paper citations.

For fixing errors that can be safely detected (such as a Mathematics Subject Classification (MSC) code string that is invalid in the MSC 2000 standard) procedures are formulated and coded in XSchema generated also from a web-based interface (forms). Other sets of constraint checkers run as overnight jobs together with updates of the database and metadata statistics and logs useful for the management of Metadata Editor workflow.

Finally, various detection procedures for possible errors have been suggested, evaluated and implemented for finding anomalous and suspicious content of metadata fields, with lists of warnings generated, including hyperlinks for easy checking by an operator. An important control concerns the integrity of TEX sequences in metadata to assure seamless typesetting of article cover pages in the later stages: all metadata to be typeset are exported in one big file with unique references to the article, and typeset by XeLaTeX to check the TEX control sequences used in the metadata fields. This ensures that all of the TEX encoded mathematics converts into MathML format smoothly. Similar procedures allow for an efficient and economical increase of metadata completeness and quality.

4.2 Mathematical Document Classification and Categorization

Article full texts have many applications, e.g. for document classification and categorization. Fine document classification allows document filtering to reach higher precision in information retrieval systems such as DML. The most commonly used classification system today is the Mathematics Subject Classification (MSC) scheme (www.ams.org/msc/), We have developed an MSC classifier (guessed MSC) that is able to assign top-level MSC for retro-digitized articles. Our results convincingly demonstrated the feasibility of a machine learning approach to the classification of mathematical papers [14].

Another round of experiments was done with mathematical document similarity computation. We have collected corpus of full texts of more than 40,000 articles (from DML-CZ and NUMDAM)

and we have computed paper similarities using *tfidf* [16] and Latent Semantic Analysis (LSA) [3] and Random Projection methods. Methods use a Vector Space Model, first converting articles to vectors and then using the cosine of the angle between the two document vectors to assess their content similarity [8]. The difference between the methods is that while *tfidf* works directly over tokens, LSA first extracts concepts, then projects the vectors into this conceptual space where it only computes similarity.

We are now showing the links to closest document lists in DML-CZ article landing pages to get feedback from authors and readers to evaluate metrics computed in this experiment. Given that we will enrich our full text mathematical corpus significantly (with data from JSTOR, ARXIV and other sources as planned), we hope it will help to tackle plagiarism, too.

> Automating the creation of useful digital libraries—that is, digital libraries affording searchable text and reusable output—is a complicated process, whether the original library is paper-based or already available in electronic form. (Simske and Lin [17])

5 Summary, Conclusions and Acknowledgement

We have described several steps of DML-CZ workflow, as introduced and tested developed during the project development. We carried out most of the steps ourselves, to gain expertize and retain control of fine details, allowing us to plug-in new modules arising from leading edge research in the future—there are, currently, many new developments appearing and much research underway in the digitisation area. It is advisable for smaller project to outsource most of the workflow steps.

The most time consuming and costly step is metadata handling and editing, and image transformation and editing (if it cannot be automated). Bare scanning costs amount to less than 10% of the total page costs, and even less when pages can be physically cut before being used for batch scanning.

The complexity of the full digitisation workflow should not be underestimated, especially when digitising heterogeneous sources—continuous and flexible workflow adaptation is a must.

We believe that the methods, algorithms and tools developed do represent important step towards a European (EuDML) or even world-wide framework for a digital mathematics library, evolved, bottom up, from smaller "pilot" projects.

This research has been partially supported by the grant reg. no. 1ET200190513 of the Academy of Sciences of the Czech Republic, by MŠMT grants MSM0021622419 and 2C06009. The author thanks other DML-CZ colleagues for fruitful discussions that led to the design of the workflow described there and to the paper reviewers for improvement suggestions. Drawing of Figure 1 by Mirek Bartošek is acknowledged.

References

[1] Miroslav Bartošek, Petr Kovář, and Martin Šárfy. DML-CZ Metadata Editor: Content Creation System for Digital Libraries. In Sojka [21], pages 139–151. `http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml`.

[2] Pascal Chevalier. i2S DigiBook Mag, issue no. 2, July 2002. `http://ww.i2s-bookscanner.com/pdf/digibook_mag_no2.pdf`.

[3] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] DML-CZ. Digitization metadata editor. `http://sourceforge.net/projects/dme/`, 2009.

[5] Ted Dunning. Statistical identification of language. Technical Report MCCS 94-273, New Mexico State University, Computing Research Lab, 1994.

[6] Carole Goble. Curating Services and Workflows: the Good, the Bad and the Downright Ugly, 2008. Keynote presented at ECDL 2008, http://www.ecdl2008.org/keynotes/.

[7] Allyn Jackson. The Digital Mathematics Library. *Notices Am. Math. Soc.*, 50(4):918–923, 2003.

[8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[9] István Marosi and László Tóth. OCR Voting Methods for Recognizing Low Contrast Printed Documents. In *Proceedings of Second International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pages 108–115, April 2006.

[10] Tomáš Mudrák. Digitalizace matematických textů (in Czech, Digitisation of Mathematical Texts). Master's thesis, Masaryk University, Brno, Faculty of Informatics, April 2006. https://is.muni.cz/th/60738/fi_m/?lang=en.

[11] Radovan Panák. Digitalizácia matematických textov (in Czech, Digitisation of Mathematical Texts). Master's thesis, Masaryk University, Brno, Faculty of Informatics, April 2006. https://is.muni.cz/th/60587/fi_m/?lang=en.

[12] Tomáš Pulkrábek. Obrazové transformace při digitalizaci textů (in Czech, Image Transformation during Digitisation). Master's thesis, Faculty of Informatics, 2008. Bachelor's Thesis Masaryk University, Brno, Faculty of Informatics, https://is.muni.cz/th/139908/fi_b/?lang=en.

[13] Yves Rangoni, Faisal Shafait, and Thomas M. Breuel. OCR Based Thresholding. In *Proceedings of MVA 2009 IAPR Conference on Machine Vision Applications*, pages 3–18, May 2009.

[14] Radim Řehůřek and Petr Sojka. Automated Classification and Categorization of Mathematical Knowledge. In Serge Autexier, John Campbell, Julio Rubio, Volker Sorge, Masakazu Suzuki, and Freek Wiedijk, editors, *Intelligent Computer Mathematics—Proceedings of 7th International Conference on Mathematical Knowledge Management MKM 2008*, volume 5144 of *Lecture Notes in Computer Science LNCS/LNAI*, pages 543–557, Berlin, Heidelberg, July 2008. Springer-Verlag.

[15] Michal Růžička. Automated Processing of TEX-typeset Articles for a Digital Library. In Sojka [21], pages 167–176. http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml.

[16] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[17] Steven J. Simske and Xiaofan Lin. Creating Digital Libraries: Content Generation and Re-Mastering. In *Proceedings of First International Workshop on Document Image Analysis for Libraries (DIAL 2004)*, page 13, 2004. http://doi.ieeecomputersociety.org/10.1109/DIAL.2004.1263235.

[18] Ray Smith, Chris Newton, and Phil Cheatle. Adaptive Thresholding for OCR: A Significant Test. Technical Report HPL-1993-22, HP Laboratories Bristol, March 1993.

[19] Petr Sojka. Towards Digital Mathematical Library: Optical Character Recognition of Mathematical Texts. In Julius Štuller and Zdenka Linková, editors, *Inteligentní modely, algoritmy a nástroje pro vytváření semantického webu*, pages 110–113, Prague, 2006. Ústav informatiky AV ČR.

[20] Petr Sojka. Workflow in the digital mathematics library project: How mathematics is stored and retrieved. In J. Paralič, J. Dvorský, and M. Krátký, editors, *Proceedings of Znalosti 2006*, pages 243–247. VŠB–Technická univerzita Ostrava, 2006.

[21] Petr Sojka, editor. *Towards a Digital Mathematics Library*, Birmingham, UK, July 2008. Masaryk University. http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml.

[22] Petr Sojka, Radovan Panák, and Tomáš Mudrák. Optical Character Recognition of Mathematical Texts in the DML-CZ Project. Technical report, Masaryk University, Brno, September 2006. presented at CMDE 2006 conference in Aveiro, Portugal.

[23] Petr Sojka and Radim Řehůřek. Classification of Multilingual Mathematical Papers in DML-CZ. In Petr Sojka and Aleš Horák, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing—RASLAN 2007*, pages 89–96, Karlova Studánka, Czech Republic, December 2007. Masaryk University.

[24] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori. INFTY — An integrated OCR system for mathematical documents. In C. Vanoirbeek, C. Roisin, and E. Munson, editors, *Proceedings of ACM Symposium on Document Engineering 2003*, pages 95–104, Grenoble, France, 2003. ACM.