From Pixels and Minds to the Mathematical Knowledge in a Digital Library

Petr Sojka and Jiří Rákosník

[1] Masaryk University, Faculty of Informatics, Brno, Czech Republic
E-mail: sojka@fi.muni.cz
[2] Institute of Mathematics AS CR, Praha, Czech Republic
E-mail: rakosnik@math.cas.cz

Abstract. Experience in setting up a workflow from scanned images of mathematical papers 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 description of all main production steps. DML-CZ has recently been launched to public with more than 100,000 digitized pages.

Key words: digital mathematical library, retro-digitization, DML-CZ, digitization process

1 Motivation

The Czech Mathematical Society has closely followed initiatives worldwide aimed at the digitization of the mathematical literature. When the Academy of Sciences of the Czech Republic opened the R&D programme *Information Society* in 2004, an opportunity was set up in which a group of mathematicians, computer scientists and librarians were brought together and provided with funds for creating the Czech Digital Mathematics Library DML-CZ [1,2]. The aim of the project approved for the five years period 2005–2009 is to digitize the relevant mathematical literature published in the Czech lands. It should comprise periodicals, selected monographs and conference proceedings from the nineteenth century up until currently produced mathematical publications.

Despite the hitherto unsuccessful attempts to obtain funding for European or world-wide DML projects we hope that the DML-CZ can represent a step towards a European or even world-wide platform for a digital mathematics library, bottom-up evolved from smaller "pilot" projects.

We have been facing the problem that except of a rather general and brief position papers of the Committee on Electronic Information and Communication of the International Mathematical Union there are no clear guidelines to follow. We take inspiration mainly from the French initiatives NUMDAM [3] and CEDRAM.

In the first part of the project we concentrated on the retrodigitization of journals published in the pre-electronic era: we started from scanned *pixels*

Petr Sojka (editor): DML 2008, Towards Digital Mathematics Library, pp. 17–27.
ISBN 978-80-210-4658-0

on the page. Now we are turning to the "retro-born-digital period" (there is already an electronic form of a paper available in some format) and the modern born-digital publication processes that start as early as in authors' *minds*. The project should result in an authentic user-friendly fully validated digital library rich in content and metadata.

Our attention has been concentrated on the research journals *Czechoslovak Mathematical Journal*, *Commentationes Mathematicae Universitatis Carolinae*, *Applications of Mathematics*, *Archivum Mathematicum*, *Časopis pro pěstování matematiky a fysiky* (with all its later mutations up to the recent *Mathematica Bohemica*), *Kybernetika* and several more titles. We are going to include the Slovak journal *Mathematica Slovaca* and several journals devoted to mathematics education. The first digitized monographs comprise work by the famous Czech mathematician Bernard Bolzano.

Table 1 shows the current state of the DML-CZ data processing as numbers of pages processed.

The organization of the paper follows the workflow of the project, adapted to different types of acquired input data:

full digitization from prints: work starts from a paper copy;
full digitization 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.

We comment on both general and detailed aspects of all steps involved, as were developed throughout the course of the project realization. A reader interested in the technical details of the problem areas solved within the project is referred to see [4,5,6,7].

Section 2 describes image acquisition, scanning and operations done at the level of page bitmap images. Section 3 evaluates the state-of-the-art possibilities for getting textual representation of digitized papers. The process of editing metadata is discussed in Section 4. Specific handling of documents from the "retro-born-digital" period is described in Section 5. The pilot project aiming at instant import of new articles with validated metadata from journal publishers is described in Section 6. Section 7 is devoted to a generation of the final PDF files and to their import into the digital library for presentation. We finally survey our future plans in Section 8.

2 On the Pixel Level

Processing of scanned images is aimed at final delivery of 600 DPI bi-tonal images suitable for a quality OCR and a fine print. This is the quality used for example by JSTOR and NUMDAM. Images from Göttingen and images scanned in the Digitization Centre of the Library AS CR prior to the project

Table 1. Statistics of documents digitized within DML-CZ project as of March 2008. In addition to documents scanned at the Digitization Centre (DC) of the Library of the Academy of Sciences, Prague, we have got some pages already scanned from the Göttingen Digitization Centre (GDZ).

		GDZ	DC (Jenštejn)		Metadata Editor (Brno)			
Title	Years	#scan	#scan	OCR	issues	pages	articles	refs
Czech. Math. J.	1951–1991		28,546	27,796	164	27,464	2,298	15,583
Apl. Mat./Appl. Math., Praha	1956–1993		20,222	20,222	227	19,695	1,799	10,135
Arch. Math., Brno	1965–1991	6,550	0	6,552	104	6,550	665	4,608
Commentat. Math. Univ. Carol.	1960–1990	21,430	344		125	21,550	1,901	3,231
Kybernetika	1965–1997		21,185	21,152	227	21,370	1,771	
Čas. Pěst. Mat. Fys.	1872–1950	33,779			348	33,996	3,884	
Čas. Pěst. Mat.	1951–1990	19,186			160	19,186	2,119	
Math. Bohemica	1991–2000	1,836				4,684	405	
Acta Univ. Palacki. Olomuc., Fac. Rerum Nat., Math.	1973–1982		8,806	8,749		1,266		
Acta Math. Inform. Univ. Ostrav.	1993–2006		1,468	1,450				
Proceedings			5,888	5,888		5,327	779	
Monographs			8,837	8,837		4,695		
Total		82,781	95,296	100,646	1,355	165,783	15,621	33,557

DML-CZ have bi-tonal 400 DPI quality. The difference is visible, and leads to the higher OCR error rate. We do not suggest scanning in lower quality than 600 DPI.

We perform our new scans at 600 DPI with 4-bit depth, having a space for geometrical and other transformations done on images before binarization. The primary scans are archived for a possible future reprocessing if needed. We use BookRestorer software for interactive and batch image processing in an uncompressed TIFF format. Operations performed on images are:

1. geometrical correction (narrowing the baselines and widths of the same characters on the same line);
2. cropping;
3. blur filter, 3×3 pixels, to eliminate one or two pixel size variations;
4. binarization with manually adjusted parameters for every batch (usually journal volume);
5. despeckle filter, with both white and black spotting, 3×3 pixels;
6. publish — 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 Göttingen Digitization Centre (GDZ) slightly different operations are needed as the input files are already bi-tonal and some filters are applicable only on grayscale images.

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

3 Optical Character Recognition

To have papers indexed we need to get full texts from page bitmaps via the process of optical character recognition. Also, we need to recognize logical page numbers located in every TIFF. A FineReader software development kit was used to develop a part of the Sirius system for the location and recognition of page numbers, and a batch system DML-CZ OCR [8,9] 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 one percent character error rate [9].

Among solutions and software evaluated on the plain texts the FineReader gives the best results, but it has no support for the recognition of mathematical expressions. Texts without recognized maths may be enough for a basic style indexing and search, but it is not surprising that omitting maths matters when the full texts are used for such tasks as automated text classification and categorization or computing paper similarity [10]. Therefore we evaluate the state-of-the-art possibilities for mathematical OCR.

Developers of the InftyReader system [11] gradually improve the support of European languages, MathML and LaTeX export filters and enrich the recognized database of mathematical symbols. Infty's PDF import capability is very significant to us: it is now possible 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. This would allow addition of another PDF layer with formulae in TeX notation or with other maths representation, for example.

An open question remains as to how to represent the maths for indexing. Should it be MathML, or the specially crafted term algebra allowing incorporation of structural similarity measures and term unification? Results of the project arXMLiv `http://kwarc.info/projects/arXMLiv/` for translation of arXiv `http://arxiv.org/` to XML+MathML may support the XML approach, but for the rigorous comparison dual indexing schemes should be tried in the future.

4 Handling Metadata

It is known that providing complete, correct and reliable metadata requires a very large effort. The information-rich metadata needed for a full-featured digital mathematics library comprehend more than the standard sets of metadata corresponding to the Dublin Core Metadata Element Set; the bibliographical references are of importance as well as language alternations of titles, different spellings of names, full-texts obtained by OCR and then indexed etc.

It is basically impossible to anticipate all questions and problems that may appear during a digitization of mathematical literature, especially the older one. The multilingual content of the DML-CZ makes the problem yet more complex. Every paper is provided with the original title (except for Russian ones) and with its English translation. We add additional language version of the title whenever available. This happens for example when the original paper is in Czech and there is a corresponding German entry in . We keep all these versions in the metadata.

Most mathematical journals require today putting one primary and possibly several secondary codes of the Mathematics Subject Classification (MSC) in the paper. These codes have to be assigned to the old papers, when this system was not used. This represents a particular problem while moving towards the older issues, because the terminology evolved and understanding certain expressions requires reading and understanding the whole paper. We are using machine learning techniques for the development of the primary MSC guesser [7].

The references are presented in the original languages which in fact means that the list of references for a single paper may include any combination of Czech, Slovak, English, French, Russian, German, Italian. Even though we use the developed OCR techniques with automated identification of the block of references and relatively reliable language detection, a manual work of checking and correcting still remains.

Harvesting the metadata from Zentralblatt MATH and Mathematical Reviews helps a lot but several types of problems appear: when the paper is not written in English, the English translations of titles in both databases may differ. Which one should be taken as the authority? Should the title be corrected in DML-CZ if the English translation contained in the database is not satisfactory? A similar problem concerns authors' names (ambiguous transcription), in particular Russian, Chinese and Vietnamese ones, the MSC codes etc.

Starting with *Czechoslovak Mathematical Journal* as the pilot project we have designed a workflow and developed several tools to handle the metadata at the minimum price. However, when moving on to the next journals, especially the older ones, we realized that the tools must be extended and improved.

The most important tool we use is the *Metadata Editor* (ME) [4] which has gradually developed into an efficient web application that allows the metadata editing over the Internet according to assigned structured access rights. It supports two levels of actions. On the first one the person editing the data (operator) 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 important integral part of the 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.

These functionalities in combination with remote access enable to distribute the work among several people on different levels of expertise. Hired operators (mostly students of mathematics) usually work on the first level. They inspect and correct the structure of complex objects (journal – volumes – issues – articles). Afterwards, they make the initial inspection of the metadata, add the titles in the original languages, provide notes signalizing possible problems. Experienced mathematicians then add the necessary translations, complete the missing MSC codes, provide links between related papers. They also accomplish the final revision and validation of the metadata.

We consider articles 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 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.

The task of assigning MSC codes (primary, secondary) for retrodigitized articles requires qualified mathematicians. They may be helped by MSC codes suggested by an automated classifier trained by machine learning techniques from a database of articles already classified [10,7].

For fixing errors that can be safely detected (as MSC code string invalid in MSC 2000) procedures are formulated and coded. They are automatically 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 of possible errors have been suggested, evaluated and implemented for finding anomalous and suspicious content of metadata fields, with lists of warnings generated with hyperlinks for easy checking by an operator. These procedures allow for an efficient and economical increase of metadata completeness and quality.

5 Metadata from the Retro-Born-Digital Period

There used to be periods from which journals are already available in some kind of electronic form — there is no need to scan from paper. However, such data comes in a very wide variety of formats and encodings.

From the publisher of *Archivum Mathematicum* we obtained TEX sources, the publisher of *Czechoslovak Mathematical Journal*, *Applications of Mathematics*, and *Mathematica Bohemica*) provided us with a mixture of PostScript, PDF and TEX files. Even the files for a single volume of a journal might not be homogeneous: TEX formats, macros and bibliography citation typesetting differ, so that to develop a conversion filter for every file format and every markup is not worth the effort.

We have developed several strategies for extracting reliable citation list for every paper:

- generation from BibTex file;
- starting from LATEX's thebibliography environment grabbed from TEX file massaged by Perl script;
- starting from the plain text extracted from PostScript or PDF file;
- rerunning the AMS-TEX file with modified macros that write out tagged citation externally (used for *Commentat. Math. Univ. Carolinae*).

In the case when none of the above strategies is easily applicable, we resort to the standard OCR workflow starting from an electronic version of a page.

6 Data from the Born-Digital Period

Following the idea of the project CEDRAM, we wanted to automate the import of newly published papers as much as possible. We cooperated with the editors of *Archivum Mathematicum* to develop the journal production workflow in the way that data and metadata to be imported into DML-CZ are created as an automatic byproduct of the preparation of the printed issue.

New LATEX and BibTex style files have been developed; all citations for every paper are stored in the BibTex format. A set of conversion utilities and a workflow that extensively uses the programmes `make` (a Unix tool which automates the generation and handling of dependencies), Tralics (conversion to XML), JabRef (for BibTex citation management) have been developed and the first issue of the current volume of the journal has already been prepared as a result of this pilot project. The process is described in detail in [5].

7 Digital Papers Delivery

In DML-CZ, we decided to support article (book chapter) oriented delivery usually supported in scientific digital libraries as a Springer Link, as opposed to page-oriented systems used in GDZ, for example. An article or chapter is logical deliverable unit for mathematical scholars as distinct from an audience of historians.

The generation of deliverable PDF, for every paper or book chapter consists of the following steps:

- checking that all paper metadata and digital objects were approved by the operator of the ;
- generation of a LaTeX source file with metadata for title page typesetting;
- generation of title page PDFs with PdfLaTeX. There is a full paper citation, a persistent URL and a copyright notice on the page;
- merging the title page PDF and individual PDF pages of article into one PDF;
- setting the PDF security options for viewing, printing, cut and paste etc.;
- PDF object optimization (linearization) by program `pdfopt` from GhostScript software suite;
- PDF is digitally signed using DML-CZ's certificate. This allows a recipient of a PDF to verify that it originated from the DML-CZ project. It is a standard Public Key Infrastructure (PKI) approach.
- PDF is imported into digital library, file hash is computed and the counting of download statistics begins.

7.1 IPR Issues

The intellectual property rights (IPR) issues concern the copyright of three subjects: the author, the publisher (and/or distributor) and the administrator of the digital library. Concerning the author's copyright, we have to admit that the IPR issues are not fully solved yet. The announced amendment to the Czech Copyright Act did not meet our expectations. The articulation still complies rather with the interests of commercial mainstream art neglecting the needs of science and the nature of scientific publication. Unconditional adherence to the Act would seal off the main object of the DML-CZ: to provide access to mathematical literature which has been published. The problem is that the digital copy is considered a new original version of author's work. Therefore, the Institute of Mathematics AS CR as the administrator of the DML-CZ has to face the (rather theoretical) risk of some author's request aimed at the removal of his or her digitized paper from the displayed library.

The publishers have been encouraged to make contracts with authors of paper allowing to make a digital copy of their work and to present them in the DML-CZ. The relation between the publisher and the DML-CZ is settled with a formal agreement which secures the ownership of the digital material for the publisher. Depending on the actual conditions, the publisher may have a problem with the distributor's willingness to accept a reasonable moving wall for the open access to the past journal volumes. We believe that even this issue will be met with a suitable solution in due time.

The digital library has a bit set on or off for every deliverable file, whether it may be served freely or not. The bit is implicitly set on to serve the content, but paper may be "withdrawn" from the library by setting it to off by the library administrator.

7.2 Indexing and Search, PURL, OAI-PMH

The digital library system we have chosen for running DML-CZ is the open source system DSpace developed at Boston MIT. DSpace is now actively supported and maintained by the DSpace Consortium — it has customizable layout layer Manakin, embedded OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) server and a scalable indexing library, persistent URL support for every object in the library; its complexity and maintenance is still manageable.

DSpace uses Lucene — an open source library for information retrieval (indexing and search), that can be enriched by the special handling of mathematical formulae in the future.

Discussions about a persistent URL for our PDFs ended in the semantically rich URL of the pattern `dmlcz/`*handle#-authors-title*. The mandatory part of the URL ends with the handle number, other URLs with this prefix will be redirected to the full one. Indexing robots will thus use words of the main article's metadata (author, title) with high weight for rank computation — kind of special optimizations for Google[bot].

Further details of DSpace customization for the DML-CZ web presentation `http://dml.cz` are described in [12] and [4].

8 Summary and Future Work

We have described how the digitization process is implemented in the DML-CZ project and what problems we have been facing. An overview of the highly structured and wide-ranging workflow unifying different and heterogeneous data sources shows that the complexity of building a mathematical library is usually underestimated, and complexity of the issues is neglected. Instead of waiting for black-box solutions we decided to tackle the described issues by pilot studies and perform the actual digitization task locally. We believe that this is the most straightforward way towards the envisioned EuDML or even the World Digital Mathematics Library WDML.

Hundreds of decisions from very different areas have to be made, most of them crucial to the overall project success. Some subtasks may surely be subcontracted, diminishing the expertize needed by the core project team members, but it is at the price of loosing flexibility and even quality due to the lower understanding of all details and exceptions. We preferred to use open source software and approaches that may be automated as much as possible.

In the future we plan to

- increase the coverage of DML-CZ's retro-born digital materials by developing parametrized semi-automatic conversions from PDF journal archives;
- finalize the metadata validation procedures;
- participate in defining the interfaces and conversion filters for data export for projects on European or worldwide levels;
- pursue research in the areas of mathematical document classification, indexing and retrieval, mathematical expression OCR, representation and indexing;
- re-compress images in the front layer of PDF files with JBIG2 compression as has already been available in the PDF format since Acrobat 5, and allowing significant size reduction;
- evaluate the possibility of assigning digital object identifiers (DOI) to papers in the DML-CZ, and to use the Handle System infrastructure `http://handle.net`;
- design alternative and novel user interfaces for the digital library; we are considering the graph spring layout (force-directed algorithms for graph drawing) and context techniques used in the TouchGraph project above the web of cross-referenced publications in the digital library (represented by configurable metadata views as in Visual Browser).

An official public opening of validated DML-CZ data to the community of mathematicians was launched in June 2008. Interested readers may visit `http://dml.cz` and `http://project.dml.cz`.

Acknowledgement This research has been partially supported by the grant reg. no. 1ET200190513 of the Academy of Sciences of the Czech Republic.

References

1. Sojka, P.: From Scanned Image to Knowledge Sharing. In Tochtermann, K., Maurer, H., eds.: Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management, Graz, Austria, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co. (2005) 664–672.
2. Bartošek, M., Lhoták, M., Rákosník, J., Sojka, P., Šárfy, M.: DML-CZ: The Objectives and the First Steps. In Borwein, J., Rocha, E. M., Rodrigues, J. F., eds.: Communicating Mathematics in the Digital Era. A. K. Peters, MA, USA (2008) 69–79.
3. Bouche, T.: Towards a Digital Mathematics Library? In Rocha, E. M., ed.: CMDE 2006: Communicating Mathematics in the Digital Era. A. K. Peters, MA, USA (2008) 43–68.

4. Krejčíř, V.: Building Czech Digital Mathematics Library upon DSpace System (2008) In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, pp. 117–126.
5. Růžička, M.: Automated Processing of TEX-typeset Articles for a Digital Library (2008) In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, pp. 167–176.
6. Bartošek, M., Kovář, P., Šárfy, M.: DML-CZ Metadata Editor: Content Creation System for Digital Libraries (2008) In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, pp. 139–151.
7. Řehůřek, R., Sojka, P.: Automated Classification and Categorization of Mathematical Knowledge, Springer-Verlag (2008) 15 pp. Accepted for publication in LNCS proceedings of CICM 2008 conferences.
8. Sojka, P.: Towards Digital Mathematical Library: Optical Character Recognition of Mathematical Texts. In Štuller, J., Linková, Z., eds.: Inteligentní modely, algoritmy a nástroje pro vytváření semantického webu, Prague, Ústav informatiky AV ČR (2006) 110–113.
9. Sojka, P., Panák, R., Mudrák, T.: Optical Character Recognition of Mathematical Texts in the DML-CZ Project. Technical report, Masaryk University, Brno (2006) presented at CMDE 2006 conference in Aveiro, Portugal.
10. Sojka, P., Řehůřek, R.: Classification of Multilingual Mathematical Papers in DML-CZ. In Sojka, P., Horák, A., eds.: Proceedings of Recent Advances in Slavonic Natural Language Processing—RASLAN 2007, Karlova Studánka, Czech Republic, Masaryk University, Brno (2007) 89–96.
11. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: INFTY — An integrated OCR system for mathematical documents. In Vanoirbeek, C., Roisin, C., Munson, E., eds.: Proceedings of ACM Symposium on Document Engineering 2003, Grenoble, France, ACM (2003) 95–104.
12. Bartošek, M., Krejčíř, V.: Jak se dělá digitální matematická knihovna (in Czech). In: Proceedings of AKP 2007, Liberec, Czech Republic (2007) `http://dml.muni.cz/docs/akp2007-sbornik.pdf`.