An Experience with Building Digital Open Access Repository DML-CZ

Petr Sojka

sojka@fi.muni.cz

Faculty of Informatics, Masaryk University, Brno, Czech Republic

Abstract: A succesfully built institutional or community repository (e.g. set of workflows) needs a *coordinated effort of librarians, IT specialists and representatives of users – content specialists*.

We will explain and discuss *design, technical a political decisions* behind building the Czech Digital Mathematics Library DML-CZ (`http://dml.cz`) in the context of other succesfull thematical community projects (PubMed Central, ADS, SCOAP3 and planned EuDML).

A framework developed for handling different types of mathematical publications is presented. It integrates workflow for the articles scanned from a paper, for documents from retro-born digital period (data were available in some type of electronic form) and for born-digital papers (newly published data from publishers). Experience gained, lessons learned and tools prepared during development of the Czech Digital Mathematics Library DML-CZ are described.

We describe problems of *migration of existing workflows (born-digital, retro-digital) into the repository'*. negotiations with Google Scholar towards better visibility, indexing and search, and problems of copyright and sustainability issues we have faced.

Keywords: digital mathematical library, mathematical knowledge representation, workflow, retrodigitization, DML-CZ

1 Motivation

Digital Library (DL) 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 mirroring file repositories more subtle solutions have to be devised: data curatorship changes to workflow curatorshipi and *services*.

World and it's DLs are becoming global. Some methods (e.g. citation ranking) start to work only as part of global, world-wide system. On the other hand, focus of search in [global] DLs inevitable has to support narrowing and semantic filtering for the needs of specific communities. It is the case also for the mother of sciences, *mathematics*.

There are communities and systems that start to dominate in some thematic areas: PubMed Central (PMC) is one of such system in medical domain, speeding up the 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, authors enjoy global topical ontology-based search. Researchers send their papers only to journals available in PMC as this leverages their citation indexes.

Domain of mathematical publications has not yet reached similar stage, although there are referative databases like MATHEMATICAL REVIEWS or ZBL. These databases contain additional independent reviews, but they do not have full texts of articles and thus loosing most of today's possibilities stemming from full-text availability. Another problem with mathematical publications is that they often contain many formulae that are hard to optically recognize and standard DL systems do not support their proper handling on [full-]text level. Sofar no significant initial funding for Worldwide Mathematical DL (WDML) was succesfull, leaving the floor open for 'bottom-up' smaller initiatives and projects as NUMDAM/CEDRAM/CEDRICS [3], EUCLID, JAHRBUCH, RUSDML [15], ARXIV or Czech Digital Mathematics Library (DML-CZ) [12, 2].

In this paper we describe a framework developed for handling different types of mathematical publications in the project DML-CZ. It integrates workflow for the articles scanned from a paper, for documents from retro-born digital period (data were available in some type of electronic form) and born-digital (newly published data from publishers). We report on the experience gained, lessons learned and tools prepared during the development of the DML-CZ project.

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 comprises peri-

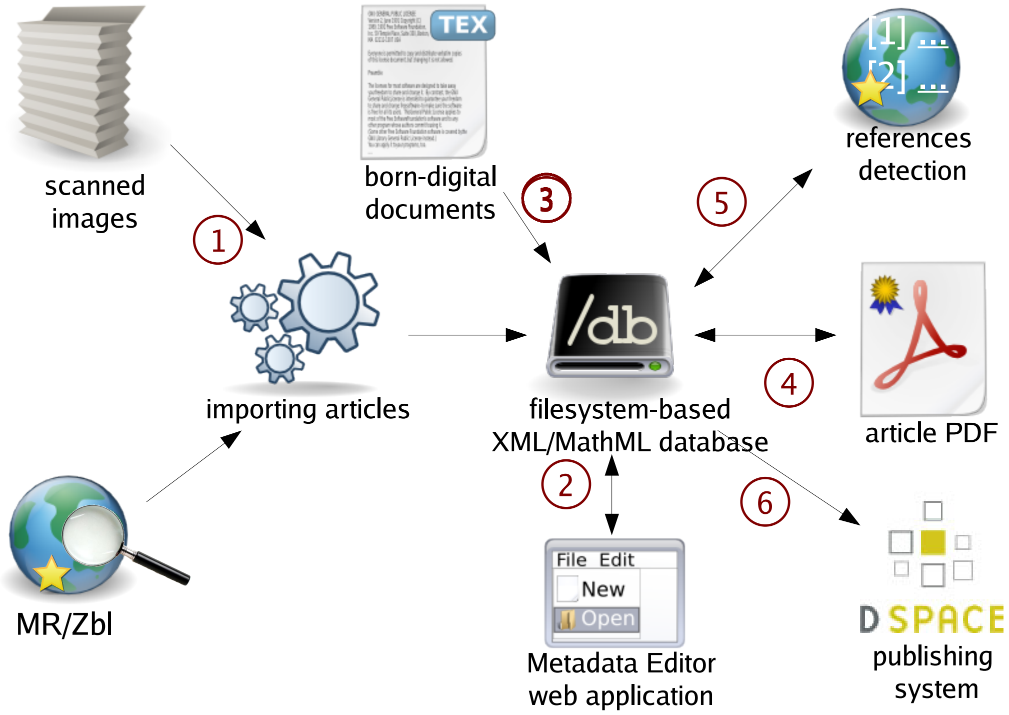

Figure 1: DML-CZ top-level workflow scheme

odicals, selected monographs and conference proceedings from the nineteenth century up until currently produced mathematical publications. It has been launched and is available on dml.cz, ready to serve 200,000 pages this year. It runs customized version of DSPACE system with adapted MANAKIN interface [7].

The general workflow of the project, shown on Figure 1 reflects 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.

Within the project, several general purpose tools have been developed, in addition to the DSpace adaptations:

- DML-CZ OCR workflow allowing recognition of scanned mathematical documents,
- web-based Metadata Editor [1],
- tools for classification of mathematical documents and measuring their similarity [9];
- workflow for born-digital publication production with direct export of metadata for DML [10] and
- plenty of other smaller tools like: extensions to Lucene engine allowing indexing of mathematics, batch PDF stamper for digitally 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 file article PDF generation with titlepage by XeLaTeX.

In the following sections we describe main features of these tools and technologies with the hope that they can be used by similar projects in other domains.

2 DML-CZ OCR

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 [14] had impressive results. We have communicated to the authors of Infty Project the possibility to combine the programmes, and got a version of the programme that is able to read PDF with a text layer inserted by FineReader [6].

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:

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 higher than three is created.

2. The computed histogram of the text block is compared 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 PDF with text layer.

4. PDF is passed to InftyReader and results are stored in Infty Markup Language (IML).

5. 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 managed to decrease the character error rate from initial 11.35% (universal language setup of FineReader) to an average 0.98% character error rate. The whole processing is fully automated after initial training. Error rate may be decreased further when Infty's character database is semiautomatically enriched when processing a new journal.

3 Metadata Editor

Metadata Editor (ME) [1, 5] has gradually developed into an efficient web application that allows simultaneous distant 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.

These functionalities in combination with remote access enable to distribute the work among several people on different levels of expertise. GUI allows hired operators (mostly students of mathematics) intuitive work on the entry 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 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 (as MSC code string invalid in MSC 2000) procedures are formulated and coded in XSchema generated also from a developed 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 of 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 a seamless typesetting of article cover pages in the later stage: all metadata to be typeset are exported in one big file with unique reference 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 the MathML format smoothly. Similar procedures allow for an efficient and economical increase of metadata completeness and quality.

4 Mathematical Document Classification and Categorization

Fine document classification allows document filtering to reach higher precision in the information retrieval system as DML. The most commonly used classification system today is the Mathematics Subject Classification (MSC) scheme (www.ams.org/msc/), Almost all of peer-reviewed mathematics journals use it, but as it has been adopted only in nineties old papers lack these classification tags. We have developed a 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 [9].

Another round of experiments was done with mathematical document similarity computation. We have collected corpus of more than 20,000 journal article fulltexts and we tried computing paper similarities using *tfidf* [11] and Latent Semantic Analysis (LSA) [4] and Random Projection methods. Methods 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 going to show the links to closest document lists in our DML-CZ article pages to get the feedback from authors and readers to evaluate metrics computed in this experiment. It helps to tackle plagiarism, too.

5 Unifying Metadata

Ways to acquire metadata for articles from from different periods (retro-digital, retro-born and born-digital) differ. Some journals have already volume of retro-digital and retroborn periods available in referative databases and import their initial version. For other journals we started from OCR texts and edited them in ME. Metadata editor together with set of transformations (in XSLT) and import filters is indispensable for these types of tasks and their proper timing (ordering) has to be ensured by the software developed.

Most publishers' workflow starts from properly tagged input data (well structured validated LaTeX or MathML). Their workflow could be adjusted only slightly to get proper validated metadata for DML-CZ DL directly as a side-effect of the main publishing process. We have been doing this kind of cooperation with several publishers switching to DML-CZ as their electronic publishing platform: Masaryk University Press for journal *Archivum Mathematicum* [10], Charles University for *Commentat. Math. Univ. Carol.* (CMUC) and we are working with Academy of Sciences ČR for journals *Math. Bohemica*, *Czech Mathematical Journal*, *Aplications of Mathematics* and *Kybernetika* and Palacky University for *Acta Univ. Palacki Olomouc*.

With the developed workflow for the born-digital data files are available in the library almost instantly together with the printed publication, without additional costs.

Conclusion and Acknowledgement

We believe that the DML-CZ and methods and tools described represent a step towards a European or even world-wide framework for a digital mathematics library, bottom-up evolved 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.

References

[1] Miroslav Bartošek, Petr Kovář, and Martin Šárfy. DML-CZ Metadata Editor: Content Creation System for Digital Libraries. In Sojka [13], pages 139–151.

[2] Miroslav Bartošek, Martin Lhoták, Jiří Rákosník, Petr Sojka, and Martin Šárfy. DML-CZ: The Objectives and the First Steps. In Jonathan Borwein, Eugénio M. Rocha, and José Francisco Rodrigues, editors, *CMDE 2006: Communicating Mathematics in the Digital Era*, pages 69–79. A. K. Peters, MA, USA, 2008.

[3] Thierry Bouche. Next Digital Mathematics Library. In Sojka [13], pages 3–15.

[4] 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.

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

[6] Toshihiro Kanahori and Masakazu Suzuki. Refinement of digitized documents through recognition of mathematical formulae. In *Proceedings of the 2nd International Workshop on Document Image Analysis for libraries*, pages 95–104, Lyon, France, April 2006.

[7] Vlastimil Krejčíř. Building Czech Digital Mathematics Library upon DSpace System. In Sojka [13], pages 117–126.

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

[9] 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 7^{th} 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.

[10] Michal Růžička. Automated Processing of TeX-typeset Articles for a Digital Library. In Sojka [13], pages 167–176.

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

[12] Petr Sojka. From Scanned Image to Knowledge Sharing. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management*, pages 664–672, Graz, Austria, June 2005. Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co.

[13] Petr Sojka, editor. *Towards Digital Mathematics Library—Proceedings of DML 2008*, Birmingham, UK, July 2008. Masaryk University.

[14] 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.

[15] Bernd Wegner. RusDML 2008: Current Facilities of the Core Archive of Digitized Russian Publications in Mathematics. In Sojka [13], pages 83–86.