Workflow of a Digital Mathematics Library[1]

Petr Sojka

NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic

sojka@fi.muni.cz

1. Motivation, DML-CZ Workflow

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 curatorship.

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 periodicals, 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 [4].

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.

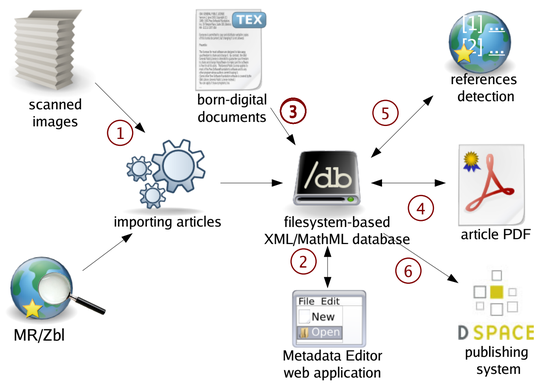

Fig. 1: DML-CZ top-level workflow scheme

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 [5];
- workflow for born-digital publication production with direct export of metadata for DML [6] 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 [9] had impressive results.

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:

(a) 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.

(b) 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.

(c) 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.

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

(e) 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.

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, 3] has gradually developed into an efficient web application that allows simultaneous distant editing according to assigned structured access rights. 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.

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 [5].

We have collected corpus of more than 20,000 journal article fulltexts and we tried computing paper similarities using *tfidf* [7] and Latent Semantic Analysis (LSA) [2] 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. 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 do 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, retroborn 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* [6], 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 this workflow for the born-digital data, files are available in the library almost instantly together with the printed publication, without additional costs.

References

[1] M. Bartošek, P. Kovář, and M. Šárfy. DML-CZ Metadata Editor: Content Creation System for Digital Libraries. In Sojka [8], pages 139–151.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

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

[4] V. Krejčíř. Building Czech Digital Mathematics Library upon DSpace System. In Sojka [8], pages 117–126.

[5] R. Řehůřek and P. Sojka. Automated Classification and Categorization of Mathematical Knowledge. In S. Autexier, J. Campbell, J. Rubio, V. Sorge, M. Suzuki, and F. 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.

[6] M. Růžička. Automated Processing of TeX-typeset Articles for a Digital Library. In Sojka [8], pages 167–176.

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

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

[9] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. 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.

European Conference on Digital Libraries, ECDL 2009, Corfu, Greece, September 27th–October 2nd, 2009

[1] Support of grants 1ET200190513 (AV ČR) and MSM0021622419 and 2C06009 (MŠMT) is acknowledged.