DML-CZ: The objectives and the first steps

Miroslav Bartošek, Martin Lhoták, Jiří Rákosník, Petr Sojka, Martin Šárfy

Encouraged by the idea of the World Digital Mathematics Library and by digitization activities worldwide, the Czech Mathematical Society initiated the digitization project "DML-CZ: Czech Digital Mathematics Library" to ensure availability of mathematical literature which has been published throughout history in the Czech lands, in digital archival form. Five specialized groups from different institutions cooperate to carry out the project. There are the first achievements as well as some questions that emerged in the course of the first eighteen months. In particular, the Metadata Editor has been developed as a specialized software to facilitate the process of creation, revision and validation of metadata obtained in an automated way.

1 Motivation

Several stimuli for the Czech mathematics digitization activity could be traced: The Czech Mathematical Society (CMS) has been involved in cooperation with Zentralblatt since 1996, when the Prague Editorial Unit was established under its auspices as one of the first external collaborating units. In 2002–2003 we skimmed digitization while contributing a little to the project ERAM by digitizing about 4 000 pages of the Jahrbuch über die Fortschritte der Mathematik. In 2003, Bernd Wegner inspired the CMS to take part in the (unfortunately, not approved) DML-EU project coordinated by the European Mathematical Society. All this experience proved useful in 2004 when the Academy of Sciences of the Czech Republic (AS CR) launched the national research and development programme Information Society. Encouraged by the idea of the World Digital Mathematics Library (WDML) [4] and by digitization activities worldwide, the CMS initiated the proposal of the digitization project "DML-CZ: Czech Digital Mathematics Library". The aim was to ensure availability of mathematical literature which has been published throughout history in the Czech lands, in digital form. The project, proposed for the period 2005–2009, has been approved [3] and has already achieved its first promising results.

2 Goals and partners

The aim of the project is to investigate, develop and apply techniques, methods and tools that will allow the creation of an infrastructure and conditions for establishing the Czech Digital Mathematics Library (DML-CZ). The mathematical literature that has been published by various institutions in the territory of the Czech lands comprises several journals of international standing, number of conference proceedings volumes, monographs, textbooks, theses and research reports. The journals include *Mathematica Bohemica* (former *Časopis pro pěstování matematiky*), *Czechoslovak Mathematical Journal* and *Applications of Mathematics* (former *Aplikace matematiky*) published by the Mathematical Institute AS CR in Prague, *Commentationes Mathematicae Universitatis Carolinae* published by the Charles University at Prague, *Archivum Mathematicum* published by the Masaryk University in Brno, *Kybernetika* published by the Institute of Information and Automation AS CR in Prague and a few others. Some of them have in the meantime been scanned in the SUB Göttingen and we plan to apply the developed procedures and tools also on this material. In view of the common history, the ongoing close cooperation and the lingual closeness of both nations of the former Czechoslovakia, we plan that the suitable Slovak mathematical literature will be included as well. Besides the retrodigitised material the existing born-digital one will be included and proper arrangements will be proposed for mostly automated incorporation of the future literature produced in electronic form. The estimated extent of the relevant literature ranges from 150 to 200 thousand pages. Upon its completion the DML-CZ will be integrated into the WDML. The techniques and tools developed for the DML-CZ might be later used for digitization in other fields of science.

The project team consists of five closely cooperating groups from different institutions, each of them being specialized in certain aspects of the digitization project. The Mathematical Institute AS CR in Prague has taken the rôle of the project coordinator and deals with selection and preparation of materials for digitization, IPR and copyright issues, and eventually will be in charge of maintenance of the developed DML-CZ system. The Institute of Computer Science of the Masaryk University in Brno is responsible for overall technical integration, development of the digital library, coordination of metadata provision and incorporation of the DML-CZ into the WDML. The Faculty of Informatics of the Masaryk University in Brno is focused on OCR processing, techniques for searching and presenting digital documents, presentation formats and relevant technology development and testing. The Faculty of Mathematics and Physics of the Charles University in Prague deals with user requirements, metadata specifications and linkage to Zentralblatt MATH and MathSciNet. This group together with the co-

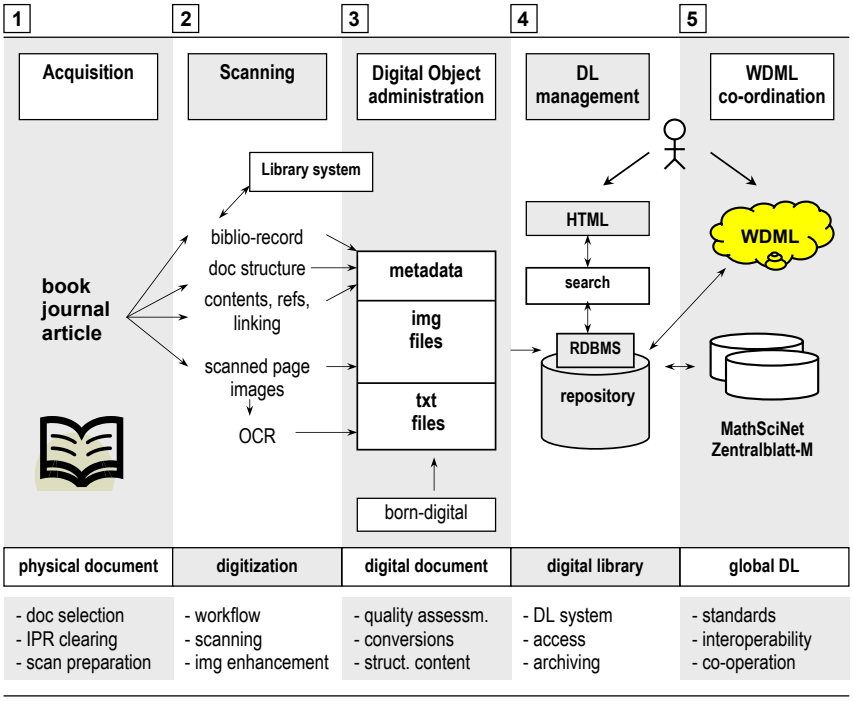

Figure 3.1. General scheme of the DML-CZ project

ordinator represent the mathematicians' view of the project. The Library AS CR in Prague ensures digitization, OCR, storage and presentation of the digitized content within the Academy of Science framework.

3 DML-CZ implementation

Do not reinvent the wheel!
[Petr Sojka]

We have analyzed [6] solutions used in other digitization efforts, especially those ongoing in Grenoble (NUMDAM) and Göttingen (DIEPER) and proposed the general scheme for the project, as depicted in Figure 3.1.

The page scanning is done in the Digitization Center of the Library AS CR.[1] With the support from the EU Solidarity Fund, the center was

[1]The Digitization Center has been established after the devastating floods in 2002 which affected a large number of library institutions in the Czech Republic.

equipped with three high-quality book scanners and the specialized software for scanned image processing. Its current production output is some 50 thousand pages a month. For the DML-CZ project two Zeutschel scanners OS 7000 (approx. 90 A4 pages per hour at 600 DPI) are used.

The test bed for the DML-CZ comprises digitized documents from the *Czechoslovak Mathematical Journal*. This journal, published since 1951, changed several times its layout as well as the publishing policy. The first two volumes were published simultaneously in Czech, Russian and multilingual versions. Beginning with the third volume only one multilingual version has been published containing papers, abstracts, book reviews and advertisements written in different languages including Czech, Slovak, Russian, English, German, French and Italian. Individual items can begin or end in the middle of pages. A typical multilingual page can be seen in Figure 3.2.

Nowadays, the content is almost exclusively in English. The papers contain free-hand drawings, graphic figures, tables and photographs. The classical typesetting was used until 1991, afterwards the journal has been typeset in TEX, the corresponding digital files are available.

The test bed consisting of almost 30 000 scanned pages has been used for the development of an automated data-flow for a complete processing of mathematical documents—starting with the scanned images and finishing with the enhanced documents (articles and interlinked metadata records) in the repository.

4 The data-flow

Creating metadata represents the most time-consuming step in digitization.
[Martin Lhoták]

4.1 Scanning. In accordance with the recommendation of the IMU Committee on Electronic Information and Communication [5] the materials are scanned in grey scale, resolution 600 (644) DPI, colour depth 4 bit in TIFF format.[2]

The Digitization Centre uses the BookRestorer software (i2S, France) for the graphical improvements of the scanned pages—mainly cropping, binarization and straightening of lines (that may be out of phase due to misalignment during digitization). The first OCR (all but mathematics) is done by the ABBYY FineReader engine integrated in the production sys-

[2] We are currently reviewing the question of optimal bit-depth for black and white document scanning: whether decreasing the colour depth to 1 bit in the scanning phase would have a significant negative impact on the OCR quality (see [7]) or not. Going down to 1 bit would speed up the workflow and lower the costs of digitization.

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*

Figure 3.2. Typical multilingual CMJ page

tem Sirius (Elsyst Engineering, Czech Republic). The system also provides an automated creation of minimal metadata with the aid of pre-defined models, the real page numbers are linked with names of graphical files.

Further improvements are aimed at finding the most suitable methods of processing the documents with a special emphasis on automated creation of metadata, the maximal OCR precision and continuing development of the production software.

The question of a long-term archiving also has to be solved yet in the project. At the moment, the backup is done using 200/400GB LTO tapes HP Ultrium II, with double copies in several stages of production. The first backup applies to TIFF files compressed by LZW, right after scanning. The second backup on tapes is done after finishing all graphical modifications and creating metadata in the production system. In the third stage, backup of the system for the end-user access is done, to ensure fast recovery of the system in emergency case. With regard to the continuously decreasing price of HDD capacity, we assume to build also the appropriate RAID array for backup.

4.2 From scanned images to articles. Design and implementation of a fully automated environment for processing mathematical documents (journal papers at the present stage) is the task of the Masaryk University team. The scheme of the developed workflow is depicted in Figure 4.3.

1. Preparation of scanned data. All the data produced in the scanning phase (page images, initial structural and page metadata) are validated with respect to their completeness and consistency, page order correctness, duplicities, etc. The data are then restructured and stored in the hierarchical Journal–Volume–Issue directory structure suitable for further processing.

2. Application of advanced OCR techniques for processing mathematical terms [10], [9]. The OCR process is realized in two phases: in the first phase the general OCR software (FineReader) is used for the text recognition with the language detection on the paragraph level (required by multilingual nature of articles in CMJ). Resulting textual layer is used for metadata autodetection and for generation of PDF files for individual pages. In the second phase the specialized OCR software (InftyReader) is applied to further elaborate the textual layer with respect to elements of mathematics (mathematical symbols, expressions, equations, etc.). The topic is treated in more detail in the paper [7].

3. Structuring the journal issue. Combination of several methods may be used to create the initial list of articles of a journal issue, and also to minimize the manual workload in this step. This includes exploitation of

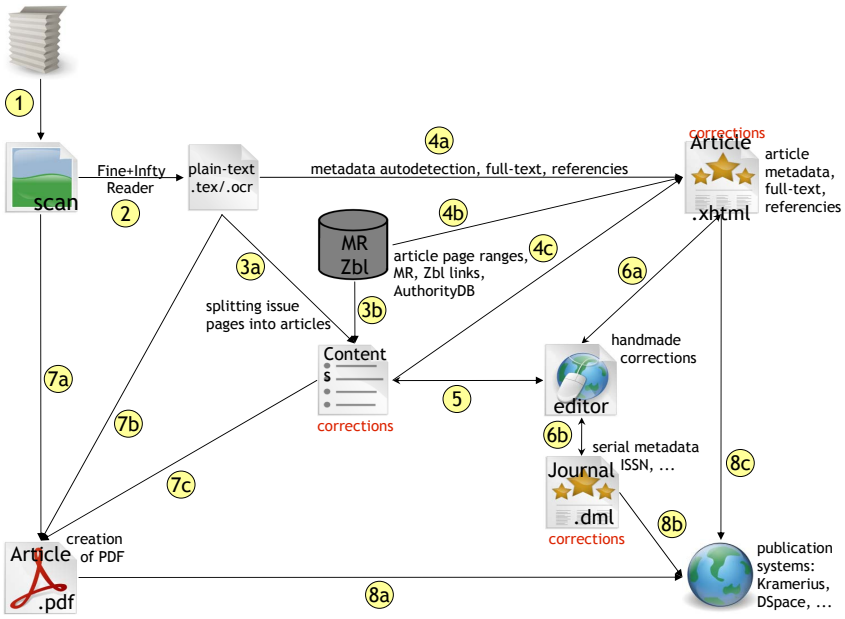

Figure 4.3. Automated workflow: From scanned images to articles

pagination information from existing databases, localization of beginnings and ends of articles in OCR-ed texts, identification of the OCR-ed Table of Contents page and of its pagination elements. Again, couple of difficulties emerged: the automatized article detection may suggest a false division of a journal issue into the parts that do not correspond to the real papers; the pagination information in databases is not always reliable, etc.

4. Autodetection of article descriptive metadata. We aimed at avoiding the manual entering of article descriptive metadata (as much as possible). The idea is to use primarily the information contained in the reference databases—in particular Zentralblatt MATH and Mathematical Reviews—to compare it and to complete it with the metadata mined from the OCR-ed text layer. Especially, in case of the older historical journal volumes, we cannot rely on the metadata from databases only and more sophisticated methods of metadata extraction have to be used.

A list of references represents the important part of article metadata. An automated identification of the block of references in the OCR-ed text is a relatively easy task; its beginning can be usually detected by localizing an appropriate keyword (References, Bibliographie, Literaturverzech-

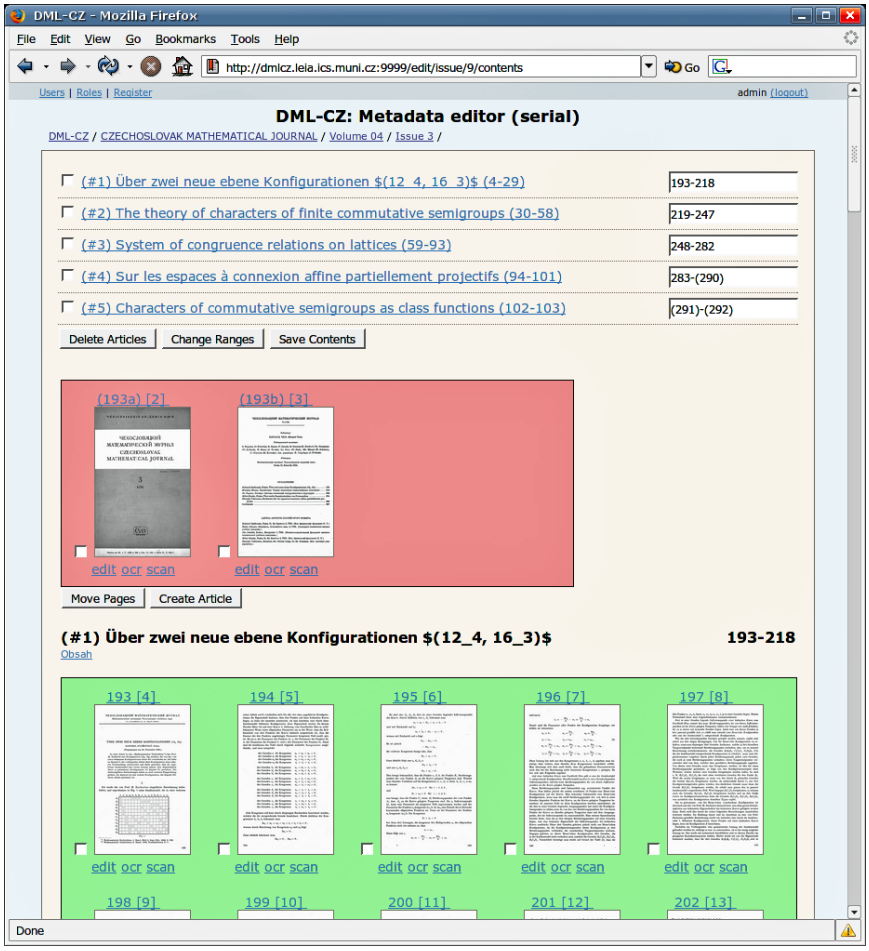

Figure 4.4. Metadata Editor: Issue structure editing

nis, Littérature, Literatura, Литература etc.) in any of the languages used in our mathematical journals. Separation of individual references and recognising their internal structure is a much more difficult problem and further research is needed to improve the quality of the process. Once a reference item is identified and structured, a linkage to reference databases can be established.

5. Manual checking of journal issue structure. The previous steps should provide as much as possible data in an automated way. However, as we have indicated, this cannot be done in an absolutely reliable way. The manual

revision and necessary corrections of data cannot be avoided and it represents a very important step in the data-flow. To make it as easy as possible the Metadata Editor has been developed. The software provides the operator with an effective support enabling a visual inspection and correction of the automatically generated structure of a journal issue (e.g. cancellation of badly identified articles and constituting the missing ones). Moreover, the Metadata Editor allows to check and revise configuration of individual articles in terms of page images assignments. This includes a visual inspection of page images content, verification of the page ordering, reshuffling pages within an article and/or between articles, removing blank pages, etc.

The Metadata Editor is designed in a very user friendly way. The operator works with the page thumbnails arranged tabularly like they were laid on the desk, as can be seen in Figure 4.4.

If any correction to journal issue structure has been implemented in this step, the process returns back to the step 4 (autodetection of article descriptive metadata).

6. Manual revision of article descriptive metadata. The Metadata Editor is also used to check and edit article metadata records. This step is important for the quality of the data stored in DML-CZ, not for the workflow itself. To avoid data inconsistencies the key metadata elements (authors' names, MSC codes) are validated against appropriate authority files in the course of editing. The validated metadata record can be collated with the appropriate scanned page which is displayed in a parallel window to ease the process of visual checking and cut&paste editing; see Figure 4.5.

The proposed metadata structure conforms, in general, to the WDML recommendations.

Several questions emerged with respect to cataloguing rules, especially in areas where unifying WDML standards are still missing—for example romanization rules for non-latin names or general WDML name authorities. For the purpose of the DML-CZ a suitable combination of the transliteration tables Zbl-new and MR-new will be used.

7. Generating article PDF files. An ultimate goal is to create the searchable PDF files combining image and textual layers for all articles. This is done automatically using the list of papers and corresponding page PDF files generated in previous steps. At the moment we have decided not to create DjVu files because the format is not so widely used among end users and its future is still unclear. Moreover, the declared advantage of DjVu's better compression had been diminished be the fact that recent versions of PDF support JBIG2 compression.

8. Exporting papers and metadata into publication systems. The Kramerius system [2] is currently used for presenting the digitized material to end

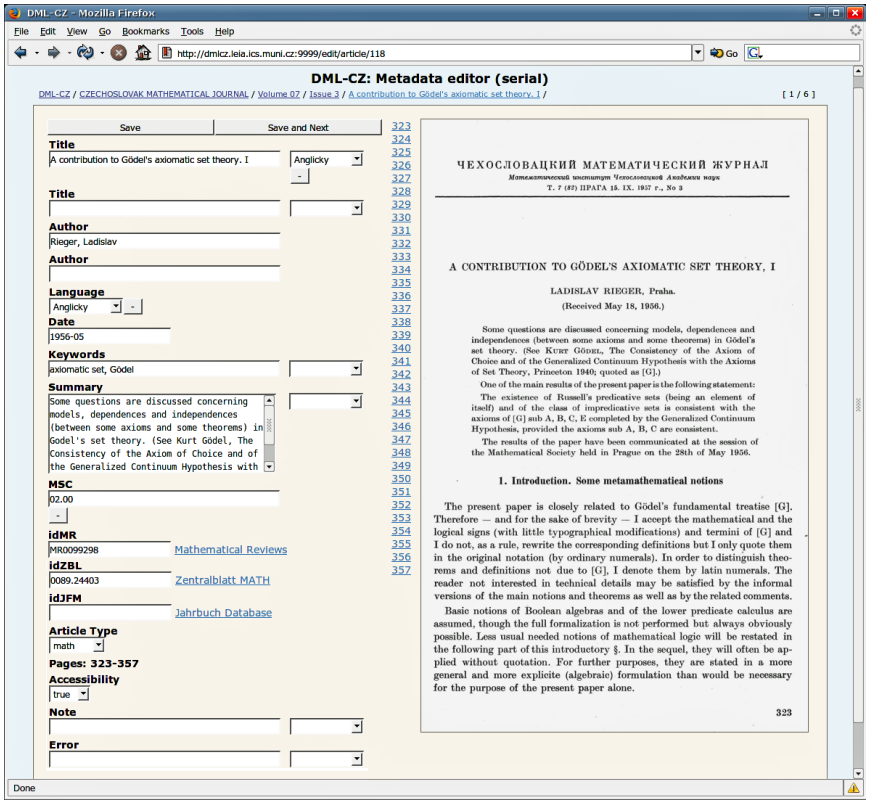

Figure 4.5. Metadata Editor: Article metadata editing

users.[3] Other digital library oriented systems or repositories, for instance the DSpace, are under consideration to host the DML-CZ as well. We shall examine also the possibility of developing a completely new presentation

[3]The Kramerius system has been developed for the Czech National Library as open source software under license GNU GPL. The Library AS CR cooperates with the Czech National Library on its enhancement. It comprises Linux, Apache web server, Apache Tomcat application server and PostgreSQL database server. XML files with metadata, together with picture files, are imported into the system. At the moment, the system supports file formats DjVu, PDF, JPG and PNG. The dynamic export of a few pages to one file is possible using PDF. The system supports the OAI-PMH protocol and we anticipate that the system will support other standards that will ensure better interoperability and persistent identification, e.g., OpenURL, PURL and DOI. Both the Library AS CR and the Czech National Library use Convera RetrievalWare engine for search within the system, which facilitates pattern search and works with Czech semantics for both metadata and full texts. To continue the development of Kramerius as a freely accessible fully-functional software, freely accessible searching tools are used as standard. It has already been used by several other Czech libraries.

system specialised for the specific needs of the DML-CZ. Whatever solution is to be used it will support the future incorporation of the DML-CZ into the WDML.

5 Conclusion and further steps

Aside from the first successful achievements there are still numerous tasks to be tackled and problems to be solved in the DML-CZ project. Some of them have already been mentioned and we can list a few more:

- *IPR issues.* Even though it is possible, in general, to rely on positive attitude of both the publishers (universities, Academy institutes, the Union of Czech Mathematicians and Physicists) and authors in granting their permissions to make the digitized document freely available, there exists certain risk following from the local legislation which is not really helpful for scientific community. We are exploring measures to minimize the limitations and to create an undepreciated digital library.

- *Further literature to be digitized.* We shall apply the achieved experience and tools developed in the test bed to incorporate further journals, conference proceedings, textbooks, theses and possibly monographs into the DML-CZ.

- *To handle the existing born-digital material* we shall design a corresponding workflow. In particular, we plan to apply the methods to enhance the files of the Czech journals digitized by the SUB Göttingen.

- *To handle the future born-digital material* we shall negotiate with the publishers and suggest arrangements in order to improve the technical quality of journals and to enable a reasonably easy incorporation of the material in the DML-CZ. The French project CEDRAM [1] may provide an inspiration.

- *Slovak journals and other literature* should be processed as well once we receive agreements of the corresponding publishers.

- *Citation references detection and linking.* We shall finish the tools for citation autodetection within individual papers. The aim is to enrich an article metadata record with a list of references linked to mathematical reference databases.

- *PDF/DjVu formats.* It is to be decided yet whether JBIG2 (about 20 kB per page) or CCITTFaxDecode compression (about 60 kB per page) will be used for the final presentation in the PDF format.

We are convinced that we can find solutions to all these problems and thus take advantage of the rare opportunity offered by the support of the DML-CZ project. In particular, we rely on cooperation with other digitization initiatives to tackle more complex problems, namely OCR of mathematics, indexation and search in mathematics, classification (completion of missing MSC codes, cf. [8]) and reference linking. On the other hand, we will be happy to share our experience and developed tools. The URL for the Czech Digital Mathematics Library is http://www.dml.cz.

Bibliography

[1] *CEDRAM: Centre de diffusion de revues académiques de mathématiques*, http://www.cedram.org.

[2] *Kramerius System. Digital Library of the Academy of Sciences of the Czech Republic*, http://kramerius.lib.cas.cz/kramerius/Welcome.do.

[3] Miroslav Bartošek, Martin Lhoták, Jiří Rákosník, Petr Sojka, and Oldřich Ulrych, *DML-CZ: Czech Digital Mathematics Library*, http://dml.muni.cz, 2005.

[4] Allyn Jackson, *The digital mathematics library*, Notices Am. Math. Soc. **50** (2003), 918–923.

[5] Committee on Electronic Information and Communication of the International Mathematical Union, *Some best practices for retrodigitization*, http://www.ceic.math.ca/Publications/retro_bestpractices.pdf.

[6] Petr Sojka, *From scanned image to knowledge sharing*, Proceedings of I-KNOW'05. Graz, Austria, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co., 2005, pp. 664–672.

[7] Petr Sojka, Radovan Panák, and Tomáš Mudrák, *Optical character recognition of mathematical texts in the DML-CZ project*, CMDE2006. Communicating Mathematics in the Digital Era. Proceedings of the Conference (E.M. Rocha, ed.), A.K. Peters, Ltd., 2006, pp. xxx–yyy.

[8] Wolfram Sperber, *Automatic classification of mathematical papers*, CMDE2006. Communicating Mathematics in the Digital Era. Proceedings of the Conference (E.M. Rocha, ed.), A.K. Peters, Ltd., 2006, pp. xxx–yyy.

[9] Masakazu Suzuki, *Refinement of digitized mathematical journals by re-recognition*, CMDE2006. Communicating Mathematics in the Digital Era. Proceedings of the Conference (E.M. Rocha, ed.), A.K. Peters, Ltd., 2006, pp. xxx–yyy.

[10] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori, *INFTY — an integrated OCR system for mathematical documents*, Proceedings of the 2003 ACM symposium on Document engineering. Grenoble, France, ACM Press, New York, NY, USA, 2003, pp. 95–104.