All for One, One for All: Several Electronic Publishing Experiences within the KT-DigiCult-Bg Project

Nikola Ikonomov

Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski Prohod St., Sofia 1113 Bulgaria
nikonomov@ibl.bas.bg

Milena Dobreva

Institute for Mathematics and Informatics, Bulgarian Academy of Sciences
8 Acad. G. Bonchev St., Sofia 1113 Bulgaria
dobreva@math.bas.bg

Abstract

The paper starts with presentation of the project "Knowledge Transfer for the Digitisation of Cultural and Scientific Heritage in Bulgaria" supported by the Marie Curie Programme of the EC and coordinated by the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, Sofia. It then describes the key areas of work in the first project year, including some faced problems and applied solutions. As a more detailed case it presents the setting for digitization of mathematical publications in Bulgaria, which should give the start of the BulDML (Bulgarian Digital Mathematical Library) project. The paper is targeted to colleagues who also work on academic projects within the field of digital preservation of and access to cultural and scientific heritage, and especially to those who work on mathematical publications.

1 Introduction

The fast development and wide application of digital methods, combined with broadened access to the Internet and falling computing costs, have created intense interest in electronic presentation and access to cultural and scientific heritage resources: original manuscripts, early printed books, epigraphic inscriptions, audio archives, immovable heritage, etc. Information and communication technologies have offered institutions new opportunities for the electronic presentation of their holdings, which are now made accessible not only to the specialists, but also to the general public worldwide.

On this setting, the electronic resources available for the Slavonic countries first of which became members of the European Union in 2004, are still scarce. This is also valid for Bulgaria which should join the EU in the year 2007. Amongst the precious holdings of Bulgarian repositories are over 12,500 manuscripts and the third largest European collection of Latin and Greek epigraphic inscriptions. Bulgarian scientific community also has an extensive range of publications in various sciences, including mathematics which still are not available electronically. However, in small countries like Bulgaria it is very difficult, because of the lack of specialists, and economically not efficient to form digitisation groups attached to the various cultural and scientific heritage institutions.

Since May 2004, the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences coordinates a Transfer of Knowledge action within the Marie Curie Program of FP6 entitled "Knowledge Transfer for the Digitisation of Cultural and Scientific Heritage in Bulgaria". The basic aim of the project is to raise to a higher level the experience in the field of digitisation and electronic publishing in the host institution and to boost practical work in various fields, including Bulgarian mathematical publications. In this project, partners of the Institute of Mathematics and Informatics in Sofia are Charles University in Prague, Czech Republic; Copenhagen University, Denmark; Trinity College Dublin, Ireland; and Institute of Informatics and Telecommunications at the National Center for Scientific Research "Demokritos", in Athens, Greece.

The basic instruments for supporting research work through this project are incoming and outgoing research visits. The outgoing visits are directed only to project partners, while incoming visits of duration not shorter than 2 months can be held in the host institution by any experienced researcher willing to share his/her experience in the fields covered by the project.

2 Basic Project Areas

The basic objective of the project is to develop the potential of the host institution (Institute of Mathematics and Informatics) in the field of digitisation of scientific heritage. The following key areas of work were defined as most important for the Bulgarian cultural and scientific heritage institutions:

- general methodology and practical setting for digitisation of cultural and scientific heritage;
- digitisation of mediaeval manuscripts (incl. digital imaging, cataloguing, text representation, electronic publishing);
- digitisation of mathematical texts and building digital mathematical libraries;
- virtual reality applications for presentation of immovable cultural heritage;
- audio archives: methods for digitisation and restoration;
- specialised quantitative methods for the study of data related to the cultural heritage;
- applications of edutainment to cultural heritage studies.

We are now at the end of our first project year and we can already report work on several directions:

- Development of the first Bulgarian electronic catalogue of Old Bulgarian manuscripts stored in Bulgarian repositories consisting of 806 records in XML format. This work was done with advice of Dr. Matthew Driscoll from Det Arnamagnæanske Institut at Copenhagen University, one of our partner institutions. The records conform to Text Encoding Initiative recommendations in this field. In this case, we have not only entered data, but created a specialised set of tools called XeditMan (XML Editor for Manuscript data) which includes an editor, visualisation tools and a component making special queries [Pavlov 2004]. Through the outgoing visit of two project members to Det Arnamagnæanske Institut, intelligent methods for search in the collection of XML records are being developed.

- For proper visualisation of Old Cyrillic texts in the descriptions (they appear in titles, textual incipits and explicits), two components are needed: good encoding system and proper font reflecting the specifics of early mediaeval Bulgarian manuscripts. Currently, the encoding of Old Cyrillic characters is not solved satisfactorily – there are numerous encoding tables associated with fonts which consist of different sets of characters. UNICODE in its Cyrillic-related part has one general table including modern and historical letters. In this direction, we initiated two activities: 1) preparing comparative sets of historical encodings and formulating suggestions for enriching UNICODE code tables, and 2) work on a new Old Cyrillic font. The font designer Filip Zrantchev on the basis of one particular manuscript, the Codex Suprasliensis, is developing this font. The aim is to offer in the electronic publications of the project a non-commercial font, which in the same time follows the current font design trends for readability both on screen and on paper.

- Work on specialised tools for improving readability of damaged texts. This is done in cooperation with our Greek partner and is based on samples from the National Archives in Bulgaria. One of the results of this cooperation is a new wavelet-based method for character recognition.

In the next three years, these activities will be continued and work in the field of digitisation of audio archives, edutainment, publications of mediaeval texts and archival records will be initiated.

Although the sources, results and ongoing work behind these topics is quite different, the project team tries to maintain active communication between the project participants involved in different tasks, so that general methodological issues might be applied – this is why we entitled our paper “All for one, one for all”. From these first endeavours we already have experiences with methods for presentation of special characters, image analysis and entering and using metadata. These will be important when we move to other subject domains.

3 Towards the BulDML project: the Bulgarian Digital Mathematical Library

In the last decade a growing group of Bulgarian specialists, mostly from the IMI, has been working on various activities related to digitisation of mathematical texts, including the full preprint process of Bulgarian mathematical journals, electronic preparation of mathematical monographs for different publishers, organizing an web-access to our journals. Recently this group took part, together with Lefkowitz & Co., in the digitisation of 16 volumes as a contribution to the ERAM project (Electronic Research Archive for Mathematics) which aim is to provide a digital archive of the most important mathematical publications of the period 1868-1942 and a database built upon the "Jahrbuch über die Fortschritte der Mathematik (1868-1943)" (Jahrbuch).
These efforts lead to the establishment and achieving good local practical experience with one particular workflow for digitisation which is described in (Kirov 2004). This workflow did not include scanning which was done by another project participant. The IMI team concentrated on use of OCR to provide partially the necessary metadata and the full text of the publications. The full text in the final electronic version was made available in T_EX.

Recently, the team started to discuss the possibilities for digitisation of the Bulgarian mathematical journals which will lead to the development of a Bulgarian Digital Mathematical Library (BulDML). This work is at its very beginning with undergoing tests of equipment and workflow design. One of the key areas of work will be connected with the use of OCR for Bulgarian mathematical texts and automatic extraction of metadata. Another relevant issue is the bilingual metadata presentation (in Bulgarian and English).

This process in Bulgaria is connected with the world efforts to build digital libraries of mathematical texts (Wegner 2002). The mathematical society worldwide is concerned with providing electronic access to publications from various countries. In the case of Bulgaria, to give some idea on the scope of the effort, we present here the basic publications and number of pages in them (see Table 1). As in all other cases, with Bulgarian publications we face the problem what should be considered as mathematics. The spring conferences of the Union of the Bulgarian Mathematicians have tracks on Informatics and a track on Education in mathematics. Amongst other publications not listed in this table but containing mathematical articles are the *Annual of Sofia University "St. Climent Ohridski" (Годишник на Софийския Университет)*, *Comptes rendus de l'Académie Bulgare des Sciences*, various publications of the Technical University and other universities in the country, monographs, archives of eminent Bulgarian mathematicians, educational journals, as well as national patents. Taking all these into account we most probably will reach over 150 000 pages to be digitised.

Title	Number of pages
Serdica	10450
Pliska	2120
Spring conferences of UBM (Доклади от Пролетните конференции на СМБ)	14456
Mathematica Balkanica	8024
Notices of the Mathematical Institute (Известия на математическия институт) – 1953-1974	3022
Physical-mathematical journal (Физико-математическо списание) – 1958-1993	10932
All together	**49004**

Table 1. List of Basic Mathematical Publications

There are several key issues which should be taken into account in the BulDML project:

- **Multilingual issues: influence on metadata and full text.** Our collection is multilingual including articles mostly in Bulgarian, French, German, Russian and English; this implies problems of proper presentation of the full texts as well as would influence the metadata structure.
- **Selection of material.** Having in mind the ongoing discussion what should be considered as Mathematics, we will probably start with purely mathematical journal for which the copyrights would be assured most quickly (from those listed in Table 1). Thus we will start with cover-to-cover digitisation. Meanwhile we will have to reach a consensus on the cases when one publication

contains articles from two or more domains. We are not aware of any ongoing effort on digitisation of scientific publications in any other field in Bulgaria.

The following basic issues are still under clarification.

1. **Clarification of cataloguing processes and metadata structure** – the underlying principle here is to assure easily the interchange with other initiatives and compliance to local Bulgarian practice in cataloguing scientific periodicals and monographs. We intend to take as a base the recommendations of the technical standards working group of the DML (Bouche, Rehmann 2003), but adaptation to local needs has still to be done. One project of particular importance is RusDML (RusDML) since it also faces bilingual metadata entry and problems with the Cyrillic metadata in the case of Russian.

   Concerning metadata we can compare the situation with the preparation of metadata for mediaeval Slavonic manuscripts. In the earlier attempts all data on manuscripts were supplied in English which made them hardly usable by the local community. For the manuscript cataloguing purposes we decided to enter data in Bulgarian and produce later a separate set of XML descriptions with English translations of the data, which will be prepared semi automatically (Pavlov 2004).

   In the case of mathematical publications, from this planning stage we envisage entering data both in the language of publication (Bulgarian, Russian, English, German, French, etc.) and in English (if the language of publication was not English). The structure of metadata is still under discussion, as well as the solution, which would assure their multilingual presentation and processing.

2. **Structure of local workflow** should assure results matching the trends in the WDML. The easiest part here relates to the quality of digital imaging processes (600 dpi of digital capture resolution in black and white/greyscale where applicable and 300 dpi for the colour sections); storage of the original scanned images in a lossless open format, most typically TIFF; including identification data in the TIFF headers). The planning of activities would require certain knowledge and this is why we started experiments with the digital capture procedures, having in mind that usually metadata require 4 times more effort compared to image capture. Any estimation on the proportion of full text preparation and linkage are not familiar to us.

3. **Production of full texts.** The Bulgarian team already has experience with the use of OCR for partial extraction of metadata and producing $\TeX$ versions of the publications. The presentation of the mathematical text in a way, which assures easy linkage to the image, is still a subject of discussion.

The work on BulDML could contribute to the following issues important for the WDML:

- Contributing to authority files for names of Bulgarian (and possibly Russian) mathematicians.

- Developing a module which would generate possible Latin transliterations of the same name written in Cyrillic. For example the name of the famous Bulgarian mathematician Н. Обрешков (1896–1963) can be discovered on the World Wide Web in the following Latin transliterations and transcriptions: Obreshkov, Obreschkov, Obreskov, Obreshkoff, Obreschkoff and even his first name has two forms – **Nikola** and **Nicola**. The forms Obreskoff and Obreškov, Obreškoff are not found on the World Wide Web but possible and should be included if one wants to perform a thorough search.

- The work on character recognition done within the KT-DigiCULT-Bg project can be applied to mathematical texts.

4 Conclusions

This paper presents in its mathematical section work, which is still on the planning stage. Since this effort is part of a greater project which deals with digital preservation of and access to scientific and cultural heritage, we hope that our future experience will contribute to the development of other Bulgarian initiatives related to digitisation of other collections of scientific publications.

We also hope to contribute through our effort to the development of the World Digital Mathematics Library.

Our project is planned with the intention to develop local best practices and guidelines. We expect feedback and advice which would help us to become an integral part of the world digital mathematics library.

References

Bouche, Rehmann (2003). T. Bouche, U. Rehmann. Digital Mathematics Library, Report of the Technical Standards Working Group. URL http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_standards_fin.pdf, , date of last visit 24 April 2005.

Jahrbuch: The Jahrbuch Project: Electronic Research Archive for Mathematics (ERAM), from URL http://ftp.gwdg.de/pub/misc/EMIS/projects/JFM/, date of last visit 24 April 2005.

Kelevedjiev E. (2005) Using Wavelets for Character Recognition. *International Journal Information Theories and Applications* (to appear).

Kirov N (2004). Standards and Technology for Digitisation of the Reference Journal "Jahrbuch", from URL http://www.math.bas.bg/~nkirov/2004/digi/std.html, date of last visit 24 April 2005 (in Bulgarian).

Pavlov P. (2004). XML Presentation of Catalogue Data on Mediæval Slavonic Manuscripts: Experience and Perspectives, In: *Proceedings of the 33rd Conference of the Union of Bulgarian Mathematicians*, Borovets, 1–4 April 2004, pp.236–240.

RusDML. RusDML: The Russian Digital Mathematics Library. URL http://www.rusdml.de/rusdml/?page=start, date of last visit 24 April 2005.

Wegner B. (2002). EMANI and Related Projects for the Long-Term Preservation of Electronic Publications. *Proc. of Online Information 2002*, p. 81-85, available on URL http://www.lmcs-online.org/Private/emani.pdf, date of last visit 24 April 2005.