МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2004
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2004
Proceedings of the Thirty Third Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 1–4, 2004

XML PRESENTATION OF CATALOGUE DATA ON MEDIAEVAL SLAVONIC MANUSCRIPTS: EXPERIENCE AND PERSPECTIVES[*]

Pavel I. Pavlov

The paper presents a brief history of preparing catalogues on mediaeval Slavonic manuscripts in electronic form. It raises two basic problems: structure of the descriptions and encoding data written in various alphabets used in the traditional catalogues. It also gives a view how the family of XML technologies can be used to produce flexible tools for encoding, manipulating and electronic publishing of catalogue descriptions.

1. Introduction. On-line presence of information is closely related to the adequate presentation of different cultures in the global information society. On the current European scene, where cultural differences and similarities are playing a key role in the integration of the Old continent, this process is of special importance. While information technologies keep offering cultural institutions a variety of opportunities for presentation and access to resources [8], this cannot be claimed for cultural heritage collections in Bulgaria. They still cannot be widely accessed in electronic form. One typical example is the mediaeval manuscript heritage.

Bulgarian repositories store about 12,500 manuscripts of Slavonic, Greek, Latin, Ottoman Turkish and other origins. Neither catalogue information on them, nor digital images could be consulted using the Internet. Although work on entering catalogue data is in the focus of interest of various research groups for about 10 years, consulting materials on local collections is still not possible.

In this paper we describe previous experience and outline some of the problems which could be addressed better with the XML technologies.

2. Previous work on electronic catalogues of mediaeval Slavonic manuscripts. Computer processing of data on mediaeval Slavonic manuscripts goes back to the 70es when database management systems were first applied to present in a structured computer form catalogue data [4]. This first attempt made clear the following problems:

1. Different specialists working in the field of mediaeval Slavonic studies have different views on the structure of the descriptions. These differences are of two basic types:

    a. *inventory of elements*—descriptions may serve different needs, and thus the number of elements may differ.

[*]This research is partially supported by the KT-Digicult-Bg project of FP6 of the EC.

   b. *subordination of elements*—various researchers have different views on the hierarchical structure of elements. For example, some may consider orthographic features in the manuscript a characteristic of a specific scribe or author of a marginal note. Some do not go into such details and consider one general orthographic description.

2. Important data which are inevitable part of the catalogue description, have to be presented in Old Bulgarian or other mediaeval languages. These data include, for example, titles, marginal notes, incipita and explicita (beginning and ending fragments in the texts). For the processing of the description, it should be made clear which data should be visualized by a specific alphabet. This is important not only for the visualization of the records, but also for their processing, e.g. to assure correct work of text search algorithms.

The first problem led to a long-lasting discussion on the structure of descriptions in the community of library specialists and researchers working on mediaeval Slavonic texts. While the library community was not happy because the solutions offered were not part of a library standard, the researchers were not happy with descriptions, which were never complete enough to cover all fields of scientific competence.

In the late 80es-early 90es of the last century, the **ISIS** library cataloguing software was introduced in Bulgaria and tested in the National Library "St. Cyril and St. Methodius". However, major cataloguing effort has not been started in that time, because of the limitations of the model. This one was matching the demand of being a library standard, but its adoption to encoding data on Slavonic manuscripts was not satisfactory.

With the advent of mark-up languages, a team in Bulgaria suggested in 1994-95 a structured description of manuscript data built as an extension of Text Encoding Initiative (TEI) [9] of that time. A project called **The Repertorium of Old Bulgarian Literature and Letters** was started as "...an archival repository capable of encoding and preserving in SGML (and, subsequently, XML) format of archeographic, palaeographic, codicological, textological, and literary-historical data concerning original and translated medieval texts represented in Balkan Cyrillic manuscripts" [6, 7]. This is an interesting example of repository project aimed to answer researchers' needs. The computer model based on SGML is discussed in [2, 3]. Currently there are 300 manuscript descriptions which should be made available on the project website (it was last checked on November 11, 2003 and there is still a message that this work is under construction).

While this was an example of a technological solution involving text mark-up, the two problems (structure of the catalogue entry and encoding mediaeval texts) stayed there.

In the late 90es, the National Library "St. Cyril and St. Methodius" and the Institute of Mathematics and Informatics joined the MASTER (**Manuscript Access through Standards for Electronic Records)** project supported by the EC [5] as associated members. It developed a TEI-conformant DTD for mediaeval manuscripts with the ambition to serve the needs of all repositories in Europe, and software for making and visualising records on manuscripts. The MASTER standard (may be with small revisions) was adopted by the TEI in May 2003.

The *Repertorium* and *Master* project are based on the same technological approach. Moreover, both projects adopted XML when it appeared. However, the models underlying both projects are different (i.e. they contain different sets of elements structured in diverse ways). Thus work of different groups in the same field needs the design of tools for migrating records from one of the encoding to the other, which is not a major technological problem. The problems with the structures of the models underlying the practical solutions are still not solved by the specialists in the subject domain.

How XML could be better applied to the field of cataloguing Slavonic manuscripts?

3. Perspectives for improving the current state. The previous experience on encoding catalogue data on mediaeval Slavonic manuscripts raised debate on the structure of the descriptions [1]. This debate should not be considered as the primary reason for the delay in the preparation and exposure of manuscript catalogues in electronic form. From information technology point of view, besides the model, we should take care for the overall process of preparation and use of the catalogue.

1. Data collection and entry. The data on manuscripts traditionally are supplied by librarians or researchers. They have different views on the details which has to be entered, as it was discussed above. This means that the model which is in circulation has to be **flexible** and **extensible**. **Migration** of data encoded following other DTDs should also be taken into account.

The practices applied so far are not assuring the best results in terms of following a certain **standard of quality of the content** (data entered). Descriptions of manuscripts may have over 100 different elements. Some of the, like illumination of the manuscript, or binding, may be described in great detail by specialists in the respective fields, but are not a subject of common interest. When a team of specialists is working on the descriptions, different participants give different quality, as it was described in [2]. A more reasonable approach would be to have a group of specialists with various competences who supply specific data (e.g., a palaeographer, a linguist, a specialist on illuminations, a specialist on bindings, a librarian familiar with the use of the manuscript). The computer tool for data entry should allow easy editing of manuscript descriptions. Functionality such as providing a list of elements, that currently are not filled in, might be of help in this process.

The tools used so far in the Bulgarian practice, the Author/Editor software in the Repertorium and NoteTab Light editor, are showing the complete structure of the manuscript description in the first case or allow including of specific elements from the description in the latter case. The first approach is somehow stressing because one starts with a huge empty template and some identifiers of the elements might be misleading. The second approach requires excellent knowledge of the description structure, which is not always the case with librarians or researchers.

We believe that a tool which allows data entry in groups of elements belonging to the same logical part of the description would facilitate the process of data entry. It also can serve better the needs of achieving common quality standard, making possible distributed data entry by a group of specialists with specific field of expertise.

2. Additional processing. Tools, which produce indices for quick reference on most specific searches, will facilitate the use of a catalogue of manuscripts in electronic form. Visualisation of single entries is not powerful enough to give quick answers on specific queries. In the beginning of the work on the Repertorium, simple and complex indices were produced be a specially designed programme tools, but in the later stages such tools are not supported. The development of a well-designed set of tools, which would organise specific data from the descriptions, would be of great help for the comparative studies of the material.

3. Data visualisation. The proper visualisation of catalogue data is of great importance for the use of the resources. The traditional approaches applied in the MASTER project are based on the visualization of the complete manuscript entry. To answer the particular needs of various target audiences, it is necessary to have a viewer, which could be tuned to the personal interests – e.g., in illumination styles, or in specific orthographic features.

4. Conclusions. Our analysis shows that in order to offer more powerful tools which would lead to faster preparation and better use of electronic resources on mediaeval manuscripts, we should develop specialised tools for data entry, processing and visualisation. A set of such tools is currently under development by the author of the article It should help to produce descriptions in electronic form faster and with better quality, and should be of great help for the study of the material by target audiences with various interests. The latter is important in the light of personalisation in the work with Internet.

REFERENCES

[1] R. Cleminson. A Good Servand but a Bad MASTER: Uses and Abuses of Standards in Manuscript Description. In: Computational Approaches to the Study of Early and Modern Slavic Languages and Texts, Proc. of the “Electronic Description and Edition of Slavic Sources” conference, 24–26 September 2002, Pomorie, Bulgaria, 105–111.

[2] M. Dobreva. Use of SGML by Philologists. In: Proceedings of SGML-Belux conference, Brussels, October 30–31 1996, 39–53.

[3] M. Dobreva. A Repertory of the Old Bulgarian Literature: Problems Concerning the Design and Use of a Computer Supported Model. In: A. Miltenova, D. Birnbaum (eds.), Medieval Slavic Manuscripts and SGML: Problems and Perspecs, Sofia, Academic Publishing House, 2000, 91–98.

[4] A. J. Geurts, A. Gruijs, J. van Krieken, W. R. Veder. Codicography and Computer, In: . V. 17–18 (1987), 4–29.

[5] MASTER, `http://www.cta.dmu.ac.uk/projects/master/`, website of the MASTER project.

[6] A. Miltenova, D. Birnbaum (Eds), Medieval Slavic Manuscripts and SGML: Problems and Perspectives, Sofia, Academic Publishing House, 2000, 372 pp.

[7] Repertorium, `http://clover.slavic.pitt.edu/~repertorium/index.html` — website of the Repertorium of Old Bulgarian Literature and Letters.

[8] S. Ross, M. Donnelly, M. Dobreva. New Technologies for the Cultural and Scientific Heritage Sector (DigiCULT, Technology Watch Report 1), ISBN 92-894-5275-7, European Commission, 2003, 196 pp.

[9] TEI, `http://www.tei-c.org/` — Text Encoding Initiative Website

Pavel Iliev Pavlov
Faculty of Mathematics and Informatics
“St. Kliment Ohridski” University of Sofia
5, James Bourchier Blvd.
1164 Sofia, Bulgaria
e-mail: `pavlovp@fmi.uni-sofia.bg`

ПРЕДСТАВЯНЕ НА ДАННИ ОТ КАТАЛОЖНИ ОПИСАНИЯ НА СРЕДНОВЕКОВНИ СЛАВЯНСКИ РЪКОПИСИ С ПОМОЩТА НА XML: ОПИТ И ПЕРСПЕКТИВИ

Павел И. Павлов

Докладът представя накратко историята на подготвянето на каталози на средновековни славянски ръкописи в електронен вид. Разглеждат се два основни нерешени проблема: структурата на описанията и въвеждането на различни азбуки, използвани в традиционните каталожни описания на ръкописи. Моделите в областта, основани на използването на XML, ще могат да се прилагат по-успешно чрез набор от програмни средства за въвеждане, обработка и визуализация на данни. Това ще ускори публикуването в електронен вид на описания на ръкописи и ще подобри начините за работа с тях, вкл. т.нар. персонализация.