Dincho Krastev
Anissava Miltenova
Andrey Boyadzhiev
Bulgarian Academy of Sciences, Sofia, Bulgaria

Sofia Corpus of Data of Slavic Manuscripts

ABSTRACT *Tlie paper is concerned with the development and exploitation of computer corpora of data from the Medieval Slavic manuscripts. The paper stresses on the ways how different information on manuscript studies could be encoded, how that encoding should be represented for interchange, and what kind of methods of searching are used. The further development of the Sofia Computerized Manuscript Corpus of Data is related with the integration of different aspects of language, text and paleography research of medieval Slavic manuscripts.*

The paper presents some results from the ongoing work and proposes new ideas for the further development of such a projects with Standard Generalized Markup Language (SGML) in the framework of the Text Encoding Initiative (TEI). The paper will focus also on standardization problems, copyright questions, teamwork organization and methodological difficulties in the development of electronic resources and their implementation in the university education.

The Central Library of Bulgarian Academy of Sciences is founded in Braile (Romania) in 1869 as a book collection of Bulgarian Learned Society. It is transferred in Sofia in 1879. In 1942 the first branch book collection is created in the "Service for; Bulgarian Dictionary" section of BAS. Since 1947 it deposits the national publishing production. In 1948 it is given a statute of Central Library of the system of the Special Libraries at the Academy.

Since 1949 it has been working with the rights of research institute in the field of library science and special bibliography. The Central Library of BAS is a multibranch scientific library responsible for the creation and organization of national library collections, information services of the researches of the Bulgarian Academy. Now the Library is well known not only in the country but abroad as a scientific center, in which modern technologies are applied.

A network of databases has been built in Central Library BAS, and a database of mediaeval Slavic manuscripts is one of very important parts of it. It is realizing one important task for librarians at the beginning of 21st century: to set up databases for ancient monuments, medieval manuscripts, early printed books and archival documents for fast access to the information they contain. The authors of this paper consider that an electronic database for the study of medieval manuscripts should cover three essential areas:

1. cataloguing of objects (manuscripts, etc.) in an adequate structure, which contains the essential data from catalogues, e.g. signature, repository, age, material, scripture, contents, bibliographic information, etc.

2. facsimiles in the form of computerized picture files, linked to the relevant entries in the catalogue database. Due to today's scanning technologies, it is possible to produce full color facsimiles of manuscripts in a satisfying quality.
3. sets of text files linked to both: the relevant catalogue database entries and the relevant manuscript facsimile picture files. These text files should provide the monument's text, encoded according to a unified transliteration standard for further processing. Even today's sophisticated OCR software packages are unable to "read" correctly manuscripts. As a result, it is still faster and more economic directly tc| type the text manually.

This three necessary elements include the possibility to support meta-information concerning specific carriers. A valuable way of is the digitization of information' for Slavic manuscripts and old printed books available from microforms, photocopies, etc. but not directly from the sources. A module, which corresponds to the description of manuscripts, is developed as an add-on to the detailed manuscript and old printed book description. Such a module enables the combination of partial codicological and text information, which is available only from microforms.

Other valuable component is bibliographical information for medieval studies. The electronic version on this part of the project began in the summer of the year 2000. Now all the relevant items on medieval Slavic languages, literature, and culture published in Bulgaria from 1990 up to the 2000 are collected and edited. The bibliography is based on the simplified XML version of the TEI idea for bibliographic reference. In this developing stage the data base consists of several units: bibliography of the books, papers, and reviews each of them linked to the information of cited works in each bibliographic item and information on the already used sources (manuscripts, old printed books, or epigraphic inscriptions). Part of this project will be published on-line in the spring of 2002, and almost all of it will be available in 2003 (as a part of the resources web pages of the Central Library of Bulgarian Academy of Sciences).

Computer-supported research and teaching in the humanities has been growing at an increasing pace over the past decades, with new methods for using computers to increase productivity in these areas. First systematic attempt to use computers in the field of Paleoslavistics took place on August 1980 at the University of Nijmegen, The Netherlands. A research team under the direction of professors A. Gruijs and C. Koster created a system for the description and cataloguing of manuscripts (Producing Codicological Catalogues with the Aid of Computers). One year later, W. Veder (Slavic Philology) joined them. At the beginning of 1984, the main components of the complimentary PCC software were available, notably the Parser, the Catalogue Print Program and Indexing Programs. At the end of 1986 J. Denev (University of Sofia) compiled the implementation of this software on IBM PC type computers (for MS-DOS) which included a Cyrillic character set and the facility to define other character sets as well. In September 1987, a conference should have taken place at the Catholic University of Nijmegen concerning the International database for Medieval Manuscripts studies, but it did not. Participants' papers were published in the journal Polata knigopisnaja, December 1987, Nr. 17-18.

Despite the expectations of paleoslavists, this theoretical platform has never been fully applied in practice. At the same time the data base system used in CIBAL (Sofia, Bulgaria) is compiled in the framework of old-fashioned ISIS library software and contains a restriction regarding the size of information in each rubric. Another disadvantage of this tool was that it was not possible to provide different types of linking between a segment in the same file and external information.

Another theoretical platform for purposes of electronic processing was given by Ljubov Dubrovina, Kiiv, Ukraine (project CODEX). Slavists also make attempts to use a number of commercial programs, which they can adjust to suit the aims of the study. Such is, for ex., the product IST, put forward and created by R. Lampe (Heidelberg, Germany). IST gives minimal possibility for coding (marking) the repertory of Slavic texts (in a limited number of margins with limitations concerning the quantity of the text and accompanied by a special code for marking the individual text units which are structured by the user). The text units (articles) are structured on the basis of a previously chosen feature, without the empirical facts from the description to be able to be compared between them. The system does not allow the usage of Cyrillic and ancient Greek fonts. The same shortcomings are typical for the system "MASTER" for cataloguing of the medieval manuscripts, experimented recently by research team of M. Driscoll (Copenhagen, Denmark).

Thus, in the field of medieval studies exists no complete coordination between Slavists and specialists in the fields of Latin, Greek and Hebrew paleography and codicology. Without an organized, technologically sophisticated initiative, the field of mediaeval Slavonic studies continued to remain isolated from modern electronic research and teaching methods. Special attention must be paid to similar editions concerning Greek and Latin medieval studies. At the same time there exists different hardware platforms as well as a wide range of software, not to mention the plethora of terminology and traditional topics of manuscript description used by specialists from different countries.

With Bulgarian-American project "Computer Supported Processing of Old Slavic Manuscripts" begun in 1994, sponsored by IREX - Washington (1994-1995) we tried to overcome this heterogeneity of approaches and ineffective attempts. A new type of software was built, which was based on the SGML (Standard Generalized Markup Language) accepted by the International Society of standardization (ISO) and especially in its TEI (Text Encoding Initiative) implementation. The goal of the project was to create a sophisticated system of processing Slavonic Manuscripts in the universal format with multiple using.

The system for computer analytical description of medieval Slavic manuscripts on the level of modern archeography, palaeography, codicology and textology was carried out from June to August 1994 in Pittsburgh, USA. The template was buit in the process of the teamwork of David Birnbaum, from the University in Pittsburgh, USA and Anissava Miltenova, Institute of literature in Bulgarian Academy of Science (IL BAS). The experiments on the program, using tests, continued almost to the end of 1994. In July-August of 1995, the last changes and specifications in the system of document type definition (DTD) were made during the visit of David Birnbaum in Sofia in the team: David Birnbaum, Anissava Miltenova, and Andrey Bojadziev. The research project was sponsored also by the foundation 'Open Society'.

The description used here is specifically intended for the developing of a Repertory of the Old Bulgarian literature and letters and is adopted for Medieval Slavic texts. The development of fonts for writing the original texts in Medieval Cyrillic belongs to research associate Rumyan Lazov from the Institute of mathematics and computing in BAS. Stanimir Velev created the searching programs a few months later. The complex description of Slavic manuscripts is built by the standard of Standard Generalized Markup Language (SGML), which was accepted by the International Society of standardization (ISO 8879). This well-known electronic standard is based on the ability to include special "markings" in the texts of natural languages, so called tags. Tagging circles certain parts of the text and signal what the data represents. It makes very easy to draw out data from the text during its computer processing. Working with Standard Generalized Markup Language (SGML) is something new to the philologists - it entered into practice in the 90's through the work of a group of specialists from France, USA, The Netherlands and Germany, who were united in the so-called Text Encoding Initiative (TEI). This standard is used here, for the first time, in the description of Medieval Slavic manuscripts and for including an arbitrary (free from limitations) sizes of non-normalized texts from the manuscripts themselves in the process of description.

The team has followed five main principles, formulated by David J. Birnbaum (http://www.slavic.pitt.edu/~djb/): 1) Standardization of document file format; 2) Multiple use (ensured by the separation of data from processing); 3) Portability of electronic texts (independence of local platforms); 4) Necessity of long-term preservation of manuscripts and archival documents in electronic form; 5) Orientation towards well-structured divisions of data according to established traditions of codicology, textology, paleography, etc.

The system for encoding of medieval Slavic text (TSM) was discussed on an international conference in Blagoevgrad six years ago (24th-28th July, 1995). The reports from the conference were published in a separate volume (Birnbaum, Boyadzhiev, Dobreva, and Miltenova 1995). The philosophy of SGML helped to settle some well known misunderstanding among paleoslavists concerning philological questions of terminology, inventory of units, character sets and data structure.

As a pilot (experimental) project-with a center at Institute of literature in BAS, over three hundred fifty manuscripts were processed by using TSM system in the SGML environment with the corresponding interface A/E *(Author/Editor,* SoftQuad, Canada) software package. Scientific papers and indices of the pilot project will be available in electronic form by Internet and in the hard copy. The book under preliminary title: "Medieval Slavic Manuscripts and SGML: Problems and Perspectives" (Sofia, 2000) is sponsored by IREX and Central European University). The articles in the book not only put into scientific circulation the achieved results from the analysis of the manuscripts, but also mark the problems that are waiting to be solved.

We consider our close prior collaboration with specialists in Slavic and general humanities computing (for ex. Institute for computational linguistics, Pisa, Italy; Portsmouth University, Great Britain, etc.) To be one of the strongest features of our both evaluative feedback on our proposals and a means of ensuring that our results will reach authoritative figures and institutions. A new stage of the project is the joint work with Prof. Ralph Cleminson on cataloguing of early printed books in Great Britain and with Dr. Martha Boyanivska (Ukraine) on description of Slavic manuscripts in the collection of National Museum in Lviv. The strong interest in joint works and exchange of electronic databases of manuscripts exists in the National library of Sankt Petersburg, Russia and in State library in Odessa, Ukraine, etc.

The Sofia project activities nowadays are concentrated on three main fields:

- The first of these is the development of the model for the processing of specifically Slavonic manuscripts and the provision, in an adequate structure, of data fields for the cataloguing of manuscripts.
- The second is the use of these principles and software to produce a database of descriptions of manuscripts in Bulgaria and, ultimately, elsewhere also (an "electronic catalogue").
- The descriptions of the manuscripts themselves constitute the first of these elements. The second will contain facsimiles in the form of computerized picture files, linked to the relevant entries in the catalogue database.
- The third main field is the development of auxiliary materials and databases. ("electronic reference books") for the study of Slavonic manuscripts, in many cases by extrapolation of the data assembled in the other phases of the project. Part of this field consists of bibliographic database for the described sources.
- As a necessary part of the manuscript description the model for digitization of microforms is developed.

These ideas have been discussed in a special panel in the framework of the 12th International Congress of Slavists, Krakow, 1998. Participants from Belorussia, Bulgaria, Chech Republic, Finland, Italy, Macedonia, Great Britain, the US, etc. put on discussion some mainstream questions in the field. One of the result from this discussion was the establishment of a special Commission to the Executive Council of the Congress for Computer Supported Processing of Slavic Manuscripts and Early Printed Books.

Part of these activities is the Master Program in the Faculty of Slavic Studies at the University of Sofia in which an essential attention is given to the knowledge in the field of markup languages, electronic transcription, text corpuses and text analysis. The Master Program includes also student training in computational linguistics and students own work on implementation of computer tools in humanities (www.slav.uni-sofia.bg/ Pages/ comhuen.htm).

We would like to emphasize, that after using portable electronic files in SGML/XML format, scientists change their point of view on the manuscripts and medieval texts, because they are going deep in the structure of medieval texts. Computer tools, using in Sofia database at the beginning of 21st century are a perspective instrument, more accurately and more comfortable for users than it was an only few years ago. Using SGML-like encoding guarantee compatibility, interchange, and multiple use of electronic editions - it is very important both for research work and for preservation of manuscripts in the libraries. We need to continue teamwork, because it is the only possible organization of such kind of project. Especially important is to joint librarians interesting in medieval manuscripts and archival documents. Of course, a strong cooperation and exchange of information in the field of computational medieval studies and in other organizations of computational humanities also is essential today and in the future. Moreover, the lack is not far to seek for the global organization dealing with the fast electronic access to the archives of different ancient cultures.

15. References

Birnbaum, D.J., A. T. Boyadzhiev, M. Dobreva, A. L. Miltenova (eds.) 1995: *Computer Processing of Medieval Slavic Manuscripts. Proceedings.* First International Conference, 24-28 July 1995, Blagoevgrad, Bulgaria. Sofia: Marin Drinov Publishing House.

Birnbaum, D. J., A. L. Miltenova (eds.) 2000: *Medieval Slavic Manuscripts and SGML: Problems and Perspectives.* Sofia: Marin Drinov Publishing House.