The Digital Library – The Bulgarian Case

Dincho Krastev
Director, Central Library
Bulgarian Academy of Sciences
1, "15 Noemvri" St., 1040 Sofia
BULGARIA

dincho@cl.bas.bg

> *…Teilhard de Chardin described the world as if he were outside of it. He was sure that every change, every new bifurcation, was going in the right direction – in the direction of increased spirituality. On the contrary, I am more impressed by the existence of multiple time horizons. A bifurcation can lead us to the best or to the worst. We are participating in an evolution whose outcome isn't clear to us. So I leave open the question of the meaning of being. I'm not even certain whether, put in these terms, a scientific answer is possible. Probably it has more to do with feelings or emotions. In any event, I believe it is more hopeful, more exhilarating, to be embedded in a living world than to be alone living in a dead universe. And this is really what I try to express in my work.*
>
> From Ilya Prigogine's interview, May 1983 by Robert B. Tucker (Omni Magazine)

1.0 INTRODUCTION

The analysis of the dynamics and the state of a wide interdisciplinary field such as library and information science (digital library including) presents various difficulties. It can be conducted from a number of viewpoints: phenomenological or historical, technological or sociological, futuristic or traditional, etc.; each of which is colored by different levels of optimism or pessimism of the respective researcher. Nevertheless, regardless of the scientific approach, the result of any study is usually presented in terms of analytical evaluation. This does not signify that the assessment of the changes is explicit. On the contrary, the evaluation process itself is more or less of an implicit character. In this regard, any serious analysis of the problems at hand should consciously take into account in relation to the current changes and dynamics of the social and technological realm.

Shortly after the end of the World War II the library science community began to discuss and ponder the future of the institution of the library in the context of the developments of information technology. A new "language" saw its dawn: that of information science. In general the term "information" is widely used in a range of fields without being strictly defined. This peculiar situation is probably due to the lack of a general consensus of what the notion of information signifies and what the subject of study of such an interdisciplinary area as information science is.

An overview of the subject of information science and its applications is so broad-ranging, limitless and fuzzy that in practice it stops short of being science in the narrow sense of the word. In fact, information science as a product of human culture is maybe the most telling modern example of the character of the human civilization as a continuous attempt to cope with the entropy understood in the widest sense.

According to these broad interpretations, information science deals with the generation, distribution, organization, retrieval and use of information, encoded in a range of ways, for numerous goals. Thus information is accessible through any *classical library* and its supporting information systems, and today through a variety of other media and channels such as Internet-based libraries, information portals and databases.

Paper presented at the RTO IMC Lecture Series on "Electronic Information Management", held in Sofia, Bulgaria, 8-10 September 2004, and published in RTO-EN-IMC-002.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **00 DEC 2004**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
The Digital Library The Bulgarian Case

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Central Library Bulgarian Academy of Sciences 1, 15 Noemvri St., 1040 Sofia BULGARIA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001735, RTO-EN-IMC-002, Electronic Information Management., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

All these factors reveal the impact of the application of information science in the last few decades on society and demonstrate the rapid expansion of the number of ways and media which provide access to information. These developments require a re-envisioning of the institution of the library and a more modern approach to library science in which the traditional limitations regarding access to collections and records, standards and formats are surmounted.

In the last decade a more modern and encompassing view of the institution of the library and the services it provides has taken hold. According to it, the library is *everything* which preserves recorded information organized according to some criteria and to which access for the public is provided. From this standpoint a library is not only a traditional library but also any bookstore, museum, individual files and, of course, the exponentially growing Internet databases and portals.

Doubtlessly, such a comprehensive interpretation of the library reflects the real interrelations between the different subjects and institutions operating in the information market. At the same time this definition is too abstract as it puts different social institutions with widely disparate missions and goals under a common denominator. This is why I prefer to think about the library as a specific type of institution possessing a being and history of its own. Furthermore, I find it more suiting for the debate to focus not on purely scientific goals but to search for more practical and direct applications.

It is worthy remembering the slightly naive general outlook which predominated in the ‘60s and ‘70s and predicted the coming disappearance not only of paper information media but also of information intermediator to which libraries then and now happily belong. The major information clients and the individual customer need, as a rule, a very well-structured and organized information, meta-information, “knowledge”. An elementary economic analysis indicates that even the largest multinational corporations do not find it profitable on a long-term basis to invest in internal departments that provide such services. Thus, in this age the principal question to be answered in the field of library science is to what extent and how fast can existing libraries adapt to this state of things and keep its role in the sphere of information provision.

It is interesting to look into the practices and operations of a number of leading institutions and structures in the informational sphere – supranational, governmental and non-governmental, academic and commercial. From the lingo they use a variety of keywords and terms which characterize the principal information processes, and directions can be deduced. Two of these notions pop up as the most salient and characteristic together with the classical notion of a library – “digital library” and “metainformation” or “metadata”.

2.0 EXAMPLES OF BULGARIAN DIGITAL LIBRARIES

The notion of the “Digital Library” is so prominent in the media (from the project for a National Digital Library presented to the President of the US and the Senate to the priorities of the 5th and 6th Framework Programmes of the European Union) that the general public does not fully realize the significance of this development. It is natural that next to the notion of the digital library lies the concept of “metainformation” or “metadata” as any library requires structured, organized databases and information which presumes the availability of a number of indexes, secondary data and analytical information retrieved and organized on a variety of meta-levels.

I will not go into the different descriptive definitions of the “Digital Library”. I will continue relying on an implicit understanding on part of the public of this concept based on its intuitive and empirical experience.

It is quite obvious and natural that the more you go to the south of Europe or South-East of Europe, the less well-structured and organized the national library and information systems you find.

Not surprisingly, not even a single project which could be characterized as the National Digital Library Initiative has been developed in Bulgaria on a national level up until now.

So, the very few modest local initiatives with relation to what could be defined as the "Digital Library" have been started by some libraries, nongovernmental and private institutions. It is worth mentioning the following examples:

- One of the most interesting, well-developed and nicely realized virtual initiatives that I could mention is the Bulgarian WebFolk.BG initiative. It has been started by a group professional from the Bulgarian Academy of Sciences (BAS) doing research in the history of music. The leader of this initiative, Professor Lubomir Kavaldjiev, is a remarkable man with a vision of how "these things" should be done. The module they have developed is dedicated to the Bulgarian authentic folk music (http://musicart.imbm.bas.bg/default-bg.htm) and it has a lot of multimedia dimensions. What is even more interesting is that it has several levels based on users' knowledge of folk music – from the highest level (for the professionals who are quite few in numbers) to the level for the general public.

- As usual, there are a lot of full-text collections (legislations, manuals, literature). The virtual full-text library of the Bulgarian authors since ancient times to the present day, "Slovo", is one of them (http://www.slovo.bg). It has been functioning successfully for about five years, covering already quite a lot of authors and texts. It is a typical example of full-text virtual collection with an emphasis to the full-text and much less to the metadata, indexes.

- An interesting joint initiative (involving the local public library and a private company) has been started several years ago in the major city of Varna. It aimed to digitize all photo and image collections (including private) of the city. Consisting of a collection of image files with a few indexes, this unique project has been functioning successfully for a number of years (http://www.libvar.bg/old-varna/index-eng.html).

- It is worth mentioning the first online Bulgarian encyclopedia, "Trud" (http://www.encyclopedia.bg). It was developed and realized by the joint efforts of a group of people from BAS and a private publishing house.

- There are also some very nicely developed virtual art galleries, including the one developed by the Central Library of the Bulgarian Academy of Sciences (http://art.cl.bas.bg/indexcl.html). Such virtual art galleries usually consist of well-designed databases with few indexes but lack three-dimensional characteristics.

- Archeologically, as is the case for all Balkan states, Bulgaria is quite a "rich" country. Yet, apart from some technological and methodological project ideas, there has been no functioning virtual, digital model of some of the significant archeological sites. There are only Web-Based info materias and one virtual tour along the corridors of an existing museum (http://www.historymuseum.org/mainset.php3?page=2). A few years ago a virtual model of the medieval Boyana church was developed thanks to the personal efforts of a single man.

- Most of the major Bulgarian mass-media publications have their online versions. Yet these online versions are not treated strictly by the national deposit law and not a single institution is officially in charge of the preservation of these publications.

I would like now to describe the activities of the Central Library of the Bulgarian Academy of Sciences (CLBAS) with regards to digitizing Slavic manuscripts. A group of researchers led by Professor Anisava Miltenova from the Institute of Literature of BAS and CLBAS have developed a most interesting project both technologically and methodologically. The project is more closely related to the metadata than to the digital objects, images themselves. It could be titled as the "Computer Supported Processing of Archival Documents and Manuscripts and their Accessibility through Communication Technologies". Let's have a closer look at what we call "SOFIA CORPUS OF DATA OF SLAVIC MANUSCRIPTS". I'll present here the experience of computer processing of Slavic manuscripts of CLBAS and ILBAS researchers.

We consider that an electronic database for the study of medieval manuscripts should cover three essential areas:

- Cataloging of objects (manuscripts, etc.) in an adequate structure, which contains the essential data from catalogs, e.g. signature, repository, age, material, scripture, contents, bibliographic information, etc.
- Facsimiles in the form of computerized image files, linked to the relevant entries in the catalog database. Scanning technologies available today make it possible to produce full color facsimiles of manuscripts in a satisfying quality.
- Sets of text files linked to both the relevant catalog database entries and the relevant manuscript facsimile image files. These text files should provide the monument's text, encoded according to a unified transliteration standard for further processing. Even today's sophisticated Optical Character Recognition (OCR) software packages are unable to "read" correctly manuscripts. As a result, it is still faster and economically more effective to type the text manually (in case you have the relevant drivers).

These three necessary elements include the possibility to support meta-information concerning specific media. One very effective and valuable outcome of the proposed approach of "digitization" of information for Slavic manuscripts and old printed books is that it could be processed from the available microforms, photocopies, without having the visual 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.

Another valuable component of the project is that of the bibliographical information for medieval studies. The electronic version of 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 Extensible Markup Language (XML) version of the Text Encoding Initiative (TEI) for bibliographic references. In this developing stage the database consists of several units including a bibliography of books, papers, and reviews, linked to cited works in each bibliographic item and to the information on the already used sources (manuscripts, old printed books, or epigraphic inscriptions).

Computer-supported research and teaching in the humanities has been growing at an increasing pace over the past decades, with new methods computer use to increase productivity in these areas. First systematic attempt to use computers in the field of Paleoslavistics took place in 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 cataloging of manuscripts (Producing Codicological Catalogues with the Aid of Computers). One year later, they were joined by W. Veder (Slavic Philology).

Historically, the coordination between Slavists and specialists in the fields of Latin, Greek and Hebrew paleography and codicology with respect to the medieval studies is far from being perfect. The field of mediaeval Slavonic studies used to be isolated from modern electronic tools and research and teaching methods for a long time. During the same period different hardware platforms and a wide range of software tools existed, along with the plethora of terminology and traditional topics of manuscript description used by specialists from different countries and schools.

With the Bulgarian-American project "Computer Supported Processing of Old Slavic Manuscripts" funded by IREX – Washington for the period of 1994–1995, we tried to overcome this heterogeneity of approaches and ineffective attempts. A new type of software was built. It was based on the Standard Generalized Markup Language (SGML), accepted by the International Standards Organization (ISO), and, especially, in its TEI implementation. This undertaking was built on the framework developed within the TEI by creating a set of modifications for manuscript description.

The major template was developed in the process of the teamwork of David Birnbaum of University in Pittsburgh, USA (http://www.slavic.pitt.edu/~djb/) and Prof. Anissava Miltenova of ILBAS, Bulgaria.

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

At this point the group of Prof. Anissava Miltenova has followed five main principles, formulated by David J. Birnbaum: 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; and 5) Orientation towards well-structured divisions of data according to established traditions of codicology, textology, paleography, etc.

The movement from a relational database management system (RDBMS) framework to SGML marked a significant reorientation in the conceptualization of computer-assisted manuscript description. More importantly, though, our SGML-based undertaking was oriented towards preparing manuscript descriptions that might be suitable for printing, electronic rendering, and searching, as was the case with the RDBMS approach.

We can observe, though, that attempts to describe Slavic manuscripts in electronic form prior to 1994 relied almost exclusively on relational or flat-file databases, an architecture that is well suited to the record-and-field nature of some bibliographic information, but that is poorly designed for representing the hierarchical structures and blocks of prose that are more natural in manuscript description.

We anticipated even at that stage (prior to 1994) that the manuscript description files would be suitable for direct analysis, so that we would be able, for example, to identify patterns of structural similarity within a corpus of manuscripts on the basis of the same raw data files that we would also use to generate traditional printed manuscript descriptions.

This database development represented important first steps in the conceptualization of Slavic manuscript description as a problem of information science, and not merely of descriptive philology and codicology, but architectural limitations inherent in the RDBMS architecture prevented these undertakings from exercising any significant, long-term influence on the practice of Slavic manuscript studies

Within the framework of the first **pilot** (experimental) project, 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 were published under the title *Medieval Slavic manuscripts and SGML: Problems and perspectives* (2000).

We consider our prior close collaboration with specialists in Slavic and general humanities computing (e.g., Institute for Computational Linguistics, Pisa, Italy; and Portsmouth University, Great Britain) to be one of the strongest features of our both evaluative feedback on our proposals and means for ensuring that our results will reach authoritative figures and institutions. A new stage of the project was the joint work with Prof. Ralph Cleminson on cataloging 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. Last year (2003) a joint contract was signed between the British Library and the CLBAS, having as a major target the processing by our group of professionals a certain collection of Slavic manuscripts from the BL. Recently a similar contract was signed with the library of the Russian Academy of Sciences in St Petersburg There is a strong interest in starting such joint projects with the National Library of St Petersburg, Russia and the State Library in Odessa, Ukraine. The same type of project is going on with Sweden (official partner Royal Academy of Sciences and as sub-partners all major Swedish libraries with Slavic manuscript collections).

The Sofia project activities nowadays are concentrated on the following 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 cataloging of manuscripts.
- Next is the use of these principles and software to produce a database of descriptions of manuscripts in Bulgaria and, ultimately, elsewhere (also an "electronic catalog").
- 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 catalog database.
- Quite an important 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 at a special panel in the framework of the 12th International Congress of Slavists, Krakow, 1998. Participants from Byelorussia, Bulgaria, Czech Republic, Finland, Italy, Macedonia, Great Britain, the US, etc. put on discussion some mainstream questions in the field. One of the results from this discussion was the establishment of a Commission to the Executive Council of the Congress for Computer Supported Processing of Slavic Manuscripts and Early Printed Books.

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

The other principal achievement of this time was the development by Stanimir Velev of a query interface for the manuscript descriptions that was prepared within the Repertorium project. Its interface was an interim solution that has now been superseded by Extensible Stylesheet Language for Transformations (XSLT) scripting, but for several years it served as the principal query engine for scholars at the Institute of Literature who were conducting philological research on the basis of our manuscript descriptions.

Our collection of articles was the first demonstration of the utility of SGML files in traditional philological research, although at that time the principal type of processing involved structured searching, a very powerful feature of SGML, but one that only scratches the proverbial surface of the capabilities of such a system.

The last phase of this process is characterized not only by the accumulation of still more manuscript descriptions, but also by the conversion of our materials from SGML to XML. The transition to XML was dictated by the remarkably broad acceptance of XML within the electronic-text community, and particularly by its adoption by the TEI, initially as an alternative to SGML, but ultimately as a replacement for it. We have currently converted over one hundred manuscript descriptions from our initial corpus of three hundred; the rest will be converted in time, and all new descriptions are being created directly in XML.

While XML was attractive because of its status as an emerging standard with a very wide following, it was also appealing because of the ancillary standards that were developed in coordination with it. In particular, we found XSLT particularly well-suited to processing XML manuscript descriptions in order to generate different views of the data, and it also provided a standards-based alternative to the database orientation that was implemented. We also used SVG, Scalable Vector Graphics, to generic graphic representations of manuscript structures. As Tommie Usdin said at the Extreme Markup 2003 conference in Montreal, "XML has made true all of the lies we told about SGML." By this she meant that SGML promised structured descriptions that could be transformed and visualized in new and different ways, but in the early days of SGML, one had to encode manuscript descriptions while taking on faith that the transformation and visualization tools would eventually be developed. XSLT and SVG have made it possible for XML to deliver the transformations and visualizations that were only foreseen, but not actually, within the early SGML context.

3.0 CONCLUSION

I would like to emphasize that, after using portable electronic files in SGML/XML format, several scientists have changed their point of view on the effectiveness of the applications of modern software tools to manuscripts and medieval texts. It is obvious how deep into the structure of medieval texts nowadays a researcher could go. Computer and software tools that are in use for the creation and maintenance of the Sofia database at the beginning of the 21st century are very powerful research instruments, more accurate and more comfortable for the users than they were only a few years ago. Using SGML/XML-like encoding guarantees compatibility, interchange, and multiple uses of electronic editions – which is very important both for research work and for preservation of manuscripts in the libraries. We need to continue the team work, because it is the only possible organization of such kind of projects. Especially important are the efforts of more libraries and archives to be involved as a common unified effort in order to preserve and make more accessible these most valuable medieval manuscripts and archival documents. Of course, a strong international cooperation and exchange of information in the field of computational medieval studies and computational humanities in general is also essential today and even more for the future.

4.0 REFERENCES AND FURTHER READING

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

Computational Approaches to the Study of Early and Modern Slavic Languages and Texts. (2003). Ed. by David Birnbaum, Anissava Miltenova, and Sarah Slevinski. Sofia.

Medieval Slavic Manuscripts and SGML: Problems and Perspectives. (2000). Ed. by Anissava Miltenova and David Birnbaum. Sofia.

Scripta & e-Scripta. The Journal of Interdisciplinary Medieval Studies. 1, 2003. Ed. by Anissava Miltenova.