Преглед НЦД 4 (2004), 46–51

Andrej Boyadžiev
(Faculty of Slavic Studies
Sofia University, Bulgaria)
andreib@slav.uni-sofia.bg

REPERTORIUM INITIATIVE. HISTORY AND PERSPECTIVES

Abstract. The last decade has witnessed systematic attempts to make access to European literary heritage easier by way of information technologies (IT). That part of the heritage which includes primary sources such as medieval manuscripts, archive documents and early printed books is an extremely demanding area of data processing, both because of the complexity of their content and because of the extensiveness of meta-data in their description, both of which have to find their adequate electronic form. If the concept of databases of descriptions or texts was prevalent before the 1990s, the method of description in the last decade was related to the application of the SGML and XML international standards. The initiative *Repertorium of Old Bulgarian Literature by Computer Means* (http://clover.slavic.pitt.edu/~repertorium) was born at the Bulgarian Academy of Sciences (BAS) in 1993-1994. The guiding principle of this initiative is that the description should service the objectives of various specialists like philologists, historians, art critics or librarians and to feature the requisite detail in every single field. As a continuation and broadening of its efforts a *Commission on Computer Processing of Medieval Manuscripts and Early Printed Books* was established in the framework of the International Committee of Slavists.

The *Repertorium of Medieval Bulgarian literature and letters* began in the year 1994 when the initial proposals and structures for describing Slavic manuscripts using SGML were first compiled by David Birnbaum (University of Pittsburgh, USA), Anisava Miltenova (Institute of Literature, BAS, Bulgaria), Milena Dobreva (IMI, BAS, Bulgaria), Berend van Dijk and Garry Gaylorg (University of Groningen). In 1995 the model was further edited with changes and specifications in the *document type definition* (DTD) by David Birnbaum, Anisava Miltenova and Andrej Bojadžiev. Since then over 300 manuscripts have been processed using this system (cf. Birnbaum et al. 1995; Miltenova, Boyadzhiev 2000: 44–55; Dobreva 2000: 91–98).

A wide variety of ideas concerning electronic data bases of Slavic manuscripts and early printed books was the main topic at a special session in 1998 at the 12th International Congress of Slavists in Kraków, where participants from 10 countries discussed crucial issues pertaining to the application of computer technology in Slavic studies. One of the results of this discussion was the establishment of a special *Commission for Computer-Supported Processing of Medieval Slavic Manuscripts and Early Printed Books* to the Executive Council of the Congress. The Commission decided to continue the *Repertorium Initiative* as a community project (Birnbaum, Miltenova 2000). The principal participants in the projects are: Institute of Literature at BAS, Faculty of Slavonic Studies at the University of Sofia, Slavic Department at the University of Pittsburgh, and Slavic Department at the University of Gothenburg.

In 2002 the members of the *Repertorium Initiative* and *Commission* members decided to move to XML technology. As a first step the SGML model was transferred

to an XML DTD set of elements with some minor corrections for the description. Then, the discussion on the model was raised again. For such an initiative in humanities computing as this is it was extremely important to renew the technological framework up to the current date and to provide some possibilities for editing the model in accordance with the experience of the team. Both technological and philological requirements needed therefore to be taken into account.

The decision to move to XML had several important consequences.

1. To write early Cyrillic without transliteration using Unicode technology. Within the Unicode framework, the Cyrillic writing system, both modern and early, could be treated as one. For example, early and modern Cyrillic may be rendered differently where appropriate by being treated as two different glyphs associated with one and the same character, so they share a single cell – Cyrillic portion of the Unicode inventory. From a rendering perspective, two approaches are possible**:**

- Using the same font for early Cyrillic as for modern Cyrillic.
- Creating separate fonts with different glyphs representing earlier forms of the same letters at the same addresses as the modern ones.

2. To represent the content of the description directly in the browsers. Modern browsers have XML parsers which are capable of validating the content of our XML files. Some of them understand the languages for the visualization of this information in the correct way prescribed by the W3 consortium in languages such as CSS or XSL, making it possible to publish the XML document electronically.

3. To include the specifications from the XML family of recommendations, such as XML linking language (XLink) and XML Path Language (XPath) for addressing parts of an XML Document, XML Pointer Language (XPointer), which extends the possibilities of this language, or Xquery for querying the XML data.

Moving to XML allowed the model to be revised from two different points of view:

1. From the encoder's perspective

2. From the editor's perspective (how the encoder and editor could control the content and appearance of this information)

The original Repertorium SGML DTD strictly divided the description into two parts: description and original texts. This division is made according to the framework of the TEI and reflected the philosophy that the description is metadata as compared to the original medieval texts. This view is reflected in the structure of the model:

```
<TEI.2>
         <teiHeader>
         ...
                  <sourceDesc>
                           <catalogueStmt>Here the
manuscript description begins ... </catalogueStmt>
                                          ...
                  </sourceDesc>
         ...
   <profileDesc>
     ...
                  <articleContentDesc id=”ACD1”>Here
follows the information on title , author, etc. given by
the researcher. </articleContentDesc>
     ...
```

```
        </teiHeader>
        <text>
                <body>
                        <div decls="ACD1">
                                <title>The original
title from the manuscript </title>
                                <inicipit></incipit>
                                <p></p>
                                <explicit></explicit>
                        </div>
                </body>
        </text>
</TEI.2>
```

It was decided to redesign the model in such a way that it could prevent mistakes in the descriptions and make the DTD structure more convenient for the encoder. There are basically two ways of doing this: putting all the information in the <teiHeader> as metadata, or regarding the description as part of the actual data, and entering into the <teiHeader> only information relating to the file creation and processing. Both views have their supporters in the markup language technology. In another major project in humanities computing in this field, "Manuscript Access through Standards for Electronic Records" (*MASTER*), the main description element <msDescription> could be entered as part of the metadata in <sourceDesc> as well as part of the <body>. The choice will probably depend on whether one wishes to make a description as part of an edition (in which case the catalogue could be seen as constituting metadata), or if the file is intended to be a holder for the description alone. The Repertorium Initiative chooses the latter variant for its model.

The other main improvement to the model was the incorporation of XML linking language (XLink) and the corresponding TEI element for links within (element <ref>) and external to (<xref>) the document. In this way it became possible, without the transformation of the file to some other format, e.g. HTML, to browse the existing files directly with some XLink sensitive browser such as DocZilla.

The editorial process was further simplified by the attachment of a CSS file to the XML document in order to view the data more easily.

In addition, the model specified further the philological data as regards paleographical and codicological features, and provided some solutions for texts with complex structure. Finally, simple linkage was made between the descriptions and images from the manuscripts.

Following discussion at the International conference "Electronic Description and Edition of Slavic Sources" held in Pomorie in 2002 (Miltenova, Birnbaum, Slevinski 2003), the team decided to include an important part from the *MASTER* model for the manuscript description in its DTD: the elements which provide historical and custodial information on the manuscript. In this way the possibility of transferring data between the two models increased.

As a result from this discussion a thorough comparison of the two approaches was made. One of the conclusions of this comparison was that the data from both models could be transferred between them without much data lose in the areas, such as:

A. catalogue and library information

B. <history> and <additional>: these elements contain the information on

the manuscript's history and custodial information.

In other areas only part of the information is comparable and could be transferred: the content of the whole manuscript and a small part from the content of each item.

The structure of the codicological and paleographic information remains the main problem which hampers the possible transfer between the two systems. This is not a matter of different element name. Here the whole philological conception differs, which is a consequence of different views in codicology and paleography. As a rule, the Repertorium model is much more detailed in the information from these areas. This means that an element with only general meaning (<p>) in MASTER would correspond to the specific elements in Repertorium.

On the other hand, there is in the MASTER model an element <msHeading> which should contain a brief structured description of a manuscript, e.g.:

```
<msHeading>
<author>Domenico Cavalca</author>
<title>Vite dei santi padri</title>
<origPlace>Naples</origPlace>
<origDate>1474</origDate>
</msHeading>
```

The idea of the *Repertorium Initiative* is that these data could be extracted from the already encoded description.

One of the principles in compiling the new Repertorium model is that the users and encoders should have the possibility of retaining views on various types of catalogues, e.g. a short catalogue or analytic description. This problem is solved by providing different ways of entering the data. For example, in many parts of the electronic description structure you could enter both: short information on a particular subject, or a structured description of the topic. This is the case in the description of decoration and orthography, where the element <overview> is provided for the short survey of the features as well as for the introduction to the subject. Such an approach could be useful especially in electronic description, which could begin as a short catalogue record and then could be further extended to a more lengthy analytic description.

One of the further steps was to test the model by working on different sorts of editions: an edition with strong stress on manuscript layout, and an edition with critical apparatus and linguistic information.

The simplest and easiest approach will include entering at first only the main features of the text, those related to text segmentation and page layout. Then, at a second stage, the electronic text can be enriched with other informational units, which will depend on the text typology. A reader selection could be then made in the framework of XML as transformation rules, or in some other manner. It could be then used in providing a main basis for the higher education in the fields such as Old Bulgarian (Old Church Slavonic), history of Slavonic languages and the history of the Medieval Slavonic texts and culture.

This approach has parallels in the preparation of corpora in linguistics. The core of a given text in a linguistic corpus could consist of only brief information connected with speaker's interaction, accents, intonation, pauses etc. This basic text could then be accompanied with lexical, morphological or syntactic information. In this way, both the edition of the medieval texts and the compilation of sources for corpus linguistics could

have similar steps in the formation of their databases.

Currently, the *Repertorium Initiative* works together with two other projects in the Central Library of the Bulgarian Academy of Sciences and the Central University Library. The main aim of the first project is to compile electronic bibliographical records for materials in Slavic Medieval Studies in Bulgaria from 1990-2000.

The other project is related to the general information on Slavonic studies. It is called "Bibliotheca Slavica" and is an electronic portal for information in the field of Slavonic Studies. It is expected that its first variant will be published electronically by the end of the current year.

Organizations and Projects

- *Bibliotheca Slavica*. To be available by the end of the year 2003. Follow the link from the University of Sofia Central University Library http://www.libsu.uni-sofia.bg/
- *Commission for Computer Supported Processing of Medieval Slavic Manuscripts and Early Printed Books* to the Executive Council of the Congress. http://clover.slavic.pitt.edu/~repertorium/commission/
- *MASTER: Manuscript Access through Standards for Electronic Records* http://www.tei-c.org.uk/Master
- Repertorium Initiative: *Repertorium of Medieval Bulgarian literature and letters.* http://clover.slavic.pitt. eu/d~repertorium/

TEI: *Text Encoding Initiative Consortium*. *Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition.* C. M. Sperberg-McQueen and Lou Burnard (eds.). 2001. http://www.tei-c.org/P4X/

Languages and Specifications

- CSS: *Cascading Stylesheet Language.* Second edition. http://www.w3.org/TR/REC-CSS2/
- HTML: *HyperText Markup Language* 4.01. http://www.w3.org/TR/HTML/
- SGML: International Organization for Standardization. Information Processing – Text and Office Systems – *Standard Generalized Markup Language* (ISO 8879), Geneva: ISO, 1986.
- Unicode: *The Unicode Standard Version 4.0.* The Unicode Consortium, Reading, Mass. 2003. http://www.unicode.org/
- XLink: *XML Linking Language.* Version 1.0. W3 Consortium Recommendation. 27 June 2001. http://www.w3.org/TR/xlink/
- XML: *Extensible Markup Language.* 1.0. Second edition. Bray, T; J. Paoli; C. M. Sperberg-McQueen; E. Maler; (eds.). W3 Consortium Recommendation, October 2000. http://www.w3.org/TR/REC-xml/
- XPath: *XML Path Language.* W3 Consortium Recommendation. 16 November 1999. http://www.w3.org/TR/xpath/

Xpointer: *XML Pointer Language.* W3 Consortium Working Draft. 16 August 2002. http://www.w3.org/TR/xptr/

References

[1] Birnbaum et al. (eds.) 1995. *Computer Processing of Medieval Slavic Manuscripts*. Proceedings. First International Conference, 27–28 July 1995, Blagoevgrad, Bulgaria. David J. Birnbaum, Andrej T. Bojadžiev, Milena Dobreva, Anisava L. Miltenova (eds.). Sofia: Marin Drinov Academic Publishing House.

[2] Birnbaum, David J., Anisava Miltenova (eds.) 2000. *Medieval Slavic Manuscripts and SGML. Problems and Perspectives*. Sofia: Marin Drinov Academic Publishing House.

[3] Dobreva, Milena 2000. *A Repertory of Old Bulgarian Literature: Problems Concerning the Design and the Use of a Computer-Supported Model*. In: Birnbaum, Miltenova 2000: 91–98.

[4] Miltenova, Anisava, David Birnbaum, Sarah Slevinski (eds.) 2003. *Computational Approaches to the Study of Early and Modern Slavic Languages and Texts.* Proceedings of the "Electronic Description and Edition of Slavic Sources" conference. 24-26 September 2002, Pomorie, Bulgaria. Sofia: Boyan Penev Publishing Center.

[5] Miltenova, Anisava, Andrej Boyadzhiev 2000. *An Electronic Repertory of Medieval Slavic Literature and Letters: A Suite of Guidelines*. In: Birnbaum, Miltenova 2000, 44-68.

Andrej Boyadžiev
Faculty of Slavic Studies,
Sofia University, Bulgaria
mailto:andreib@slav.uni-sofia.bg