XML EDITORS OVERVIEW AND THE CHALLENGE TO BUILD XML-ORIENTED EDITOR FOR MEDIAEVAL MANUSCRIPT DESCRIPTIONS[1]

Pavel Pavlov

Abstract: The paper presents an overview of XML and software tools for its use, with an emphasis on XML editors. Based on the experience of two Bulgarian projects on preparing electronic descriptions of mediaeval manuscripts from the 1990es, we define the following requirements to the editor used for manuscript cataloguing: minimum elements on the screen; arrangement of elements according to the practice in the subject domain; supplying default values whenever this is possible; supplying possible values in combo boxes whenever this is possible; and ease of data entry (in Bulgarian with possibility to enter mediaeval text fragments in Old Cyrillic). These requirements were taken into account for the development of a specialized editor, XEditMan, which is presented in the article. Currently, 200 descriptions of manuscripts are available which were entered and edited using XEditMan. The average time for data entry with the editor is about three times less than the time spent in previously used software tools in Bulgaria.

Keywords: XML, XML editors, mediaeval manuscript cataloguing, XEditMan.

Introduction

The interest to digitisation of scientific and cultural heritage has been considerably growing in the last decades. The electronic access to cultural heritage is one of the priority areas of the European Commission. The Cultural Heritage Applications Unit of the Information Society Directorate General of the European Commission promoted the priorities in the field through a document known as *The Lund Principles* which put emphasis on *making visible and accessible the digitised cultural and scientific heritage of Europe*; *coordination of efforts*; *development of a European view on policies and programmes*, as well as of *mechanisms to promote good practice in a consistent manner* [Lund Principles, 2001]. Currently, most large institutions from the cultural sector are taking measures to make their collections available online. The first step in this direction is to provide access to cataloguing information about the holdings in a specific collection. In Bulgaria, this still is not done for the manuscript collections of any repository.

One of the recognised approaches on world wide scale is to use XML to present data on manuscripts. In this paper, we first give a brief overview on XML and tools, which allow its use. Then we present the experience of two Bulgarian projects in the field of manuscript cataloguing and formulate several basic requirements to a specialised editor for entering data on mediaeval Slavonic manuscripts and present our work in this direction.

These requirements were taken into account for the development of XEditMan, an XML editor for mediaeval manuscripts. The use of the editor is illustrated. One basic advantage is higher accuracy of entered data and better time characteristics (about three times faster data input compared to previously used tools).

Overview: XML and Various Types of Tools

XML (eXtensible Markup Language) is an open standard developed by the W3C (World Wide Web Consortium). It has two interconnected applications: web presentation and data exchange. One distinguished feature of XML is that it separates the encoding of data from program logic and user interface code. This leads to platform independence and reusability of resources.

XML is based on the Standard Generalized Markup Language (SGML), an ISO standard which puts the basics of many developments in the field of electronic transmission of documents through defining tag sets forming the DTD (document type definition) which are used to mark-up the electronic text and allow easy processing [ISO,

[1] The research presented here is partially supported by the project KT-DigiCult-Bg (FP6) and by the ICT Agency in Bulgaria.

1986]. SGML was designed in 1986 and was oriented towards work with large collections of documents, not towards the Web. The DTD practice of SGML was expanded in XML in order to offer more data types and allow easy building of hyperdocuments. HTML was another (earlier than XML) successor of SGML designed for visualization of documents on the Web, but its orientation to present the document layout leads to limitations on the presentation of data for processing, not just for display.

An XML application unifies several parts stored separately: data, structure, presentation and program access.

The XML data are stored as the actual XML document (it is usually called the document instance). This document contains data and mark-up elements.

The second element is a file, which contains the rules defining the specific XML document's elements, attributes, entities, and processing instructions. In the beginning, following the SGML principles, a DTD file served this purpose. Later XML Schema specification started to be used in order to solve several shortcomings of the DTD: it is too restrictive to introduce new elements and does not offer support for a number of data types. XML Schema allows creating both simple and complex data types and associating them with new element types. Thus specialists working in various fields and preparing specific documents may define the structure of their documents with a great freedom.

XSL (Extensible Stylesheet Language) is the part, which ensures presentation. It allows one to render XML elements using formatting objects. For example, CSS (Cascading Style Sheets) outputs documents in HTML format. XSLT (XSL Transformation), outputs XML document into text, HTML, or any mark-up language.

The last component is called DOM (Document Object Model) which allows accessing data encoded in XML through programs. Thus data can be separated from the specific platform.

There are several types of XML-oriented tools. The *XML editors* are used to create structured documents. From the point of view of the user, it is important to have an easy and understandable interface for entering data. The task of collecting manuscript descriptions in XML format inevitably raises the question how the data will be entered. Other types of tools are necessary basically for the IT staff, such as software for the *creation of DTD's or XML Schemas*, parsers for *validating XML* files (applications which check the documents against the DTD or the Schema); parsers for *parsing XSLT* (they prepare XML documents for presentation as text, HTML or PDF by applying the XSLT stylesheet language). Technical staff may also need a specialised editor for speeding the *creation of XSLT stylesheets*. For the work on manuscript catalogue descriptions, most important are the editor and the parser. We provide below some explanation and examples of tools from the various categories.

Editors

Editors allow users to create and edit XML documents using the proper DTD. XML editors often provide additional functionality, for example validation of the document against the DTD or schema. To facilitate the user in his/her work, editors rely on two basic methods:

- Use of colours to distinguish elements, attributes, and text, etc. for easy reading.
- Providing clickable lists of possible elements and attributes at the current cursor point in the document. These lists usually are located in the left pane of the editor window.

Popular professional XML editors are XMetaL® 4[1], xmlspy® 2004[2], NoteTab Pro[3]. Free editors are XMLCooktop[4], Bonfire Studio 1.4[5], NoteTabLight[6], Xeena[7], Xerlin[8] etc. The illustration on Fig. 1. shows a snapshot from xmlspy® 2004.

---

[1] http://www.sq.com/, last visited on 25 April 2004.

[2] http://www.xmlspy.com, last visited on 25 April 2004.

[3] http://www.notetab.com, last visited on 25 April 2004.

[4] http://www.xmlcooktop.com/, last visited on 25 April 2004.

[5] http://www.nzworks.com/bonfire/download.asp, last visited on 25 April 2004.

[6] http://www.notetab.com, last visited on 25 April 2004.

[7] http://www.alphaworks.ibm.com/tech/xeena, last visited on 25 April 2004.

[8] http://www.xerlin.org/, last visited on 25 April 2004.

Validating Parsers

Usually, the professional XML editors contain a built-in validator. Some are internal to the editor and others use a separate piece of software.

XSLT Parsers

The XSLT parsers play the role of formatting engines. They output data most often in HTML, text, PDF. They are sometimes part of the XML editor.

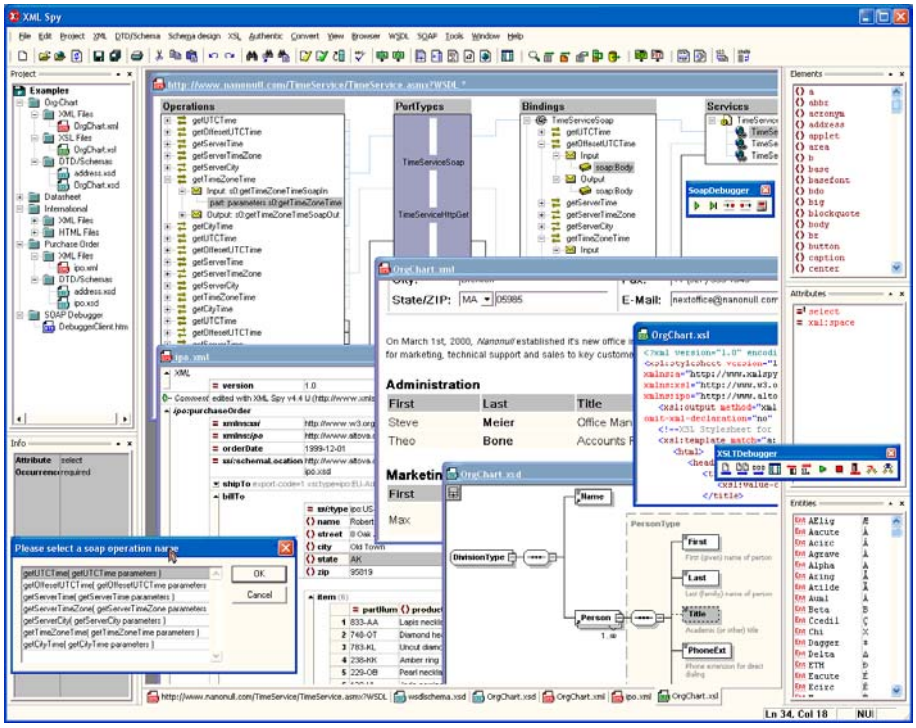

Fig. 1. Snapshot from *xmlspy® 2004, source:* http://www.xmlspy.com

The Bulgarian Experience in Preparing XML Editor for Mediaeval Manuscripts Descriptions

The idea to use a markup language for manuscript descriptions goes back to the 1990es. 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] 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, palæographic, codicological, textological, and literary-historical data concerning original and translated medieval texts represented in Balkan Cyrillic manuscripts" [Repertorium], [Miltenova et al., 2000]. This is a typical repository project aimed to answer researchers' (not librarians') needs. The computer model based on SGML is discussed in [Dobreva, 2000]. Currently there are 300 manuscript descriptions, which should be made available on the project website[1].

In the late 90es, the National Library "St. Cyril and St. Methodius" and the Institute of Mathematics and Informatics became associated members of the MASTER project *(Manuscript Access through Standards for Electronic Records)* supported by the EC [MASTER]. Within this project, a TEI-conformant DTD for mediæval manuscripts was developed with the ambition to answer 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.

[1] On April 25, 2004 there was still a message that link is disabled for file update.

In the Repertorium project, data were entered through Author/Editor software product of SoftQuad Company, a predecessor of HoTMetaL and currently available XMetaL editors. In the data entry process, users were seeing all elements from the description on the screen (surrounded by the SGML delimiters, e.g. <P> </P>) which formed long list spread on several screens. This was not very convenient, if we also add that the appearance of elements followed the structure of the DTD, which is not the same as the sequence of elements natural for the people working with mediaeval manuscripts. The organization of work was oriented towards one specialist working on one description, which produced results of different quality in the group of almost 10 specialists working on the descriptions [Dobreva, Jordanova, 2000]. The description data were entered in English which made them usable by English language speakers. To enter fragments of Old and Middle Bulgarian texts a designated font was created, and in data entry the LANG attribute was assigned to elements containing text in Old or Middle Bulgarian while for all other languages was supposed that they contain texts in English.

The experience of the pilot catalogue descriptions within the MASTER project was different in two directions: the data were entered in both Bulgarian and English with the idea that this will serve larger research community, and the editor used for the tests was NoteTabLight[1] (see Fig. 2). To enter data on both languages, elements were repeated with including of the LANG attribute showing the language of the data entered within the specific element.

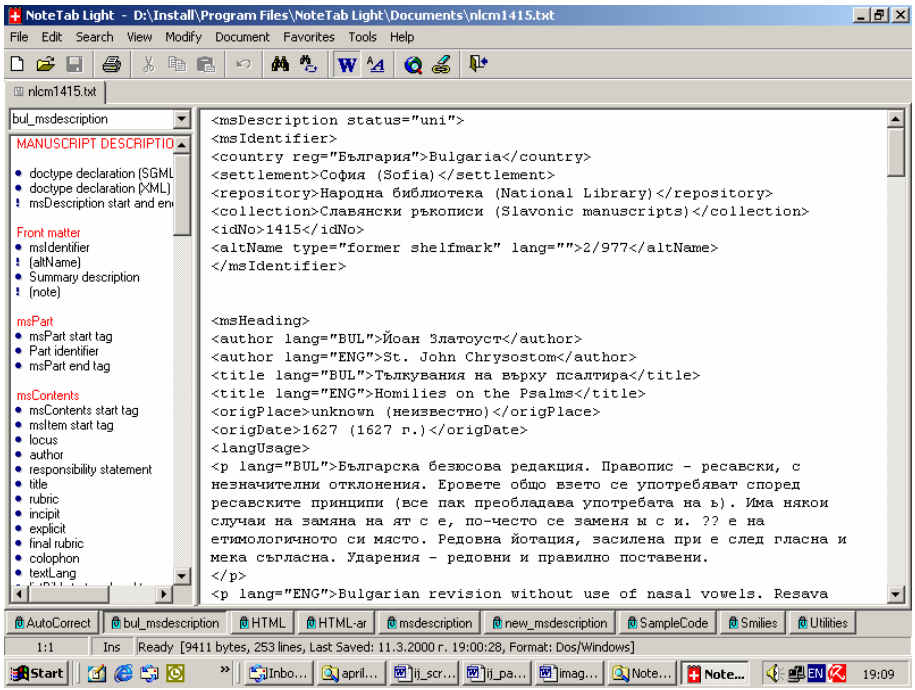

Fig. 2. Example of data entry of manuscript descriptions in NoteTabLight

Unlike the Author/Editor interface, in this case all available elements can be seen on the left pane of the window, and the person who enters data should click on the specific element which is needed. This led to high number of erroneously located elements and a heavy workload on editing the descriptions.

The experience from both projects stimulated us to formulate the following requirements to a specialized editor:

- The number of elements visible on the screen should be the minimum possible. Lengthy lists of elements confuse users who are specialists in mediaeval studies or library cataloguing who normally would enter the data. This also slows down the process of data entry and leads to mistakes.
- The sequence of appearance of the elements should follow the logic of the subject domain, not of the XML DTD.

[1] http://www.notetab.com/ntl.php, NoteTabLight – free editor offered as an alternative to the commercial professional editor NoteTab Pro.

- Quite often, the value of element is “No information” (this is because in some cases there is no information on the matter since these descriptions are based on preliminary research work on the manuscripts). To avoid multiple entry of this phrase, the value can be supplied in advance and changed by the person who enters data whenever this is needed.
- There are several elements where the values are chosen from a list: for example, names of repositories, cities, values of attributes for language, etc. To avoid errors, combo boxes with possible values could be supplied.
- Ease of entering data written in Old Cyrillic script.
- Interface in modern Bulgarian (thus specialists who enter the data see names of elements which are familiar to them, and do not have to become acquainted in details with the DTD itself.

Taking these considerations into account, we decided to create a specialized editor, which takes into account these requirements in its interface. The decision to create a home-made editor was taken after the consideration of possibilities to adapt existing commercial editors. Since the left pane with all elements listed and the alphabet encoding could not be solved satisfactorily, we decided to create a tool which could be easily installed on a computer with a running Microsoft Internet Explorer browser and Internet Information Server.

XEditMan: A XML Editor for Mediaeval Manuscripts Descriptions

XEditMan is actually a set of tools: editor for new document, editor for existing document and a visualisator. The editor is currently oriented towards the use of the MASTER DTD for manuscript descriptions adopted by TEI[1].

Data Entry: The New Document Editor and the Editor for an Existing Document

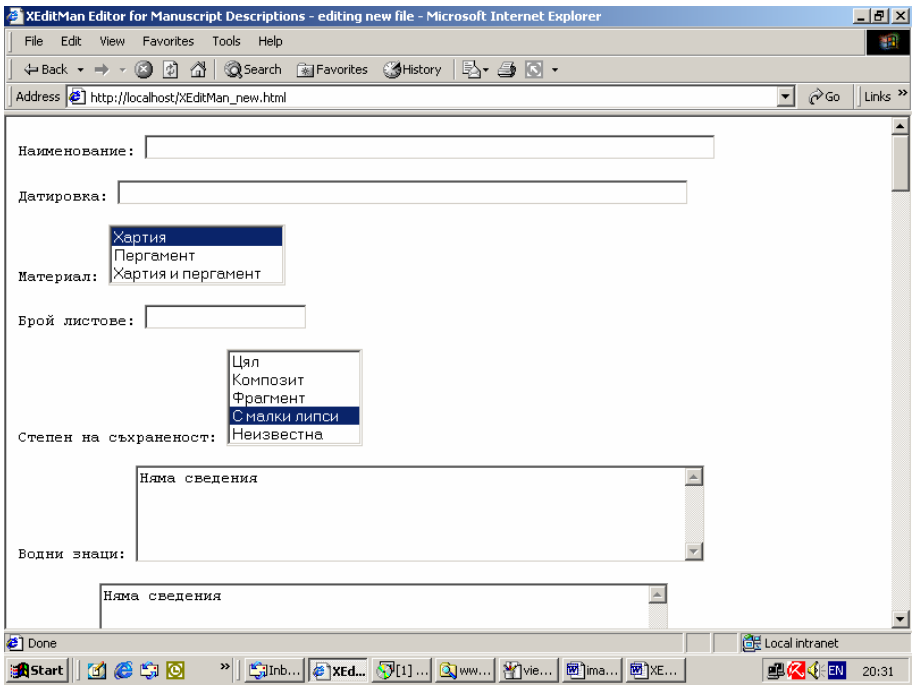

Fig. 3. XEditMan: Data entry interface

The editor of new document is used to enter data arranged in the order, which is natural for the subject domain. In two cases repetitive elements are possible: description of scribes and description of texts appearing in the manuscript. In these cases, during the first entry the user supplies the data on the first scribe (respectively, text)

[1] The relevant materials can be found on http://www.tei-c.org.uk/Master/Reference/, last accessed on April 25, 2004.

and the total number of scribes (texts). Then when the description is opened with the editor of existing texts, the respective number of elements appear in the window and make possible the entry of the information on the other scribes (respectively texts). Fig. 3 presents part of the data entry window, in which we see several types of elements: with no value; with supplied values, and combo boxes for choice of possible value.

The first two fields on Fig. 3, name and date, are typical fields for direct data entry. The third and fifth elements, material and manuscript status, are supplied with combo boxes containing possible values. In the last two elements (the visible one is a watermark) the value “No information” is entered by default. If there is no information about the element, the specialist who enters data does not have to bother with writing this text again and again.

After the data are entered, the users clicks a button “Save the description” which generates the XML document conformant to the MASTER project DTD (see Fig. 4); now all element identifiers appear according to the DTD.

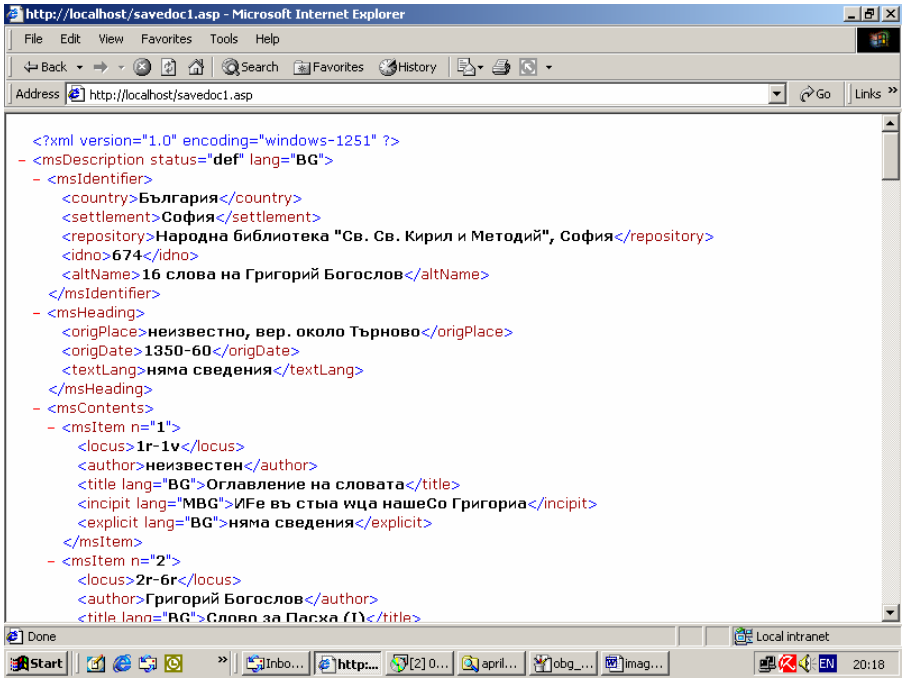

Fig. 4. A Sample of XML Document which is being saved after data were entered in XEditMan

The generation of this document is done in a way, which guarantees successful validation. This organization of work combines easy data entry and DTD-conformant result. For this reason, the editor does not include an internal validator. It is suggested to use a commercial editor for validation purposes and for cases where the interface of the editor does not support too specialized cases appearing sometimes in manuscript descriptions, like quoting within the content of specific element. We made experiments with the use in such cases of TurboXML editor (see Fig. 5). The work on XEditMan was done with the idea to cover the mass case of data entry on manuscripts. In very specific cases which appear rarely (like nesting quotes, bibliographical references and corrections to the Old Bulgarian texts), but would require too many complications in the interface, specialists who are familiar with the DTD could enter data using commercial editors which arrange the document as it is saved in XML format.

Data Visualisation

After the data are entered, they can be visualized with the help of another component of the editor. There are two modes of visualization: visualization of the complete document, which is demonstrated on Fig. 6, or visualization of selected elements from the description.

To make possible further processing of data on sets of manuscript descriptions, we are currently working on a program interface, which would extract data from XML descriptions into a database. This would provide tools for group queries.

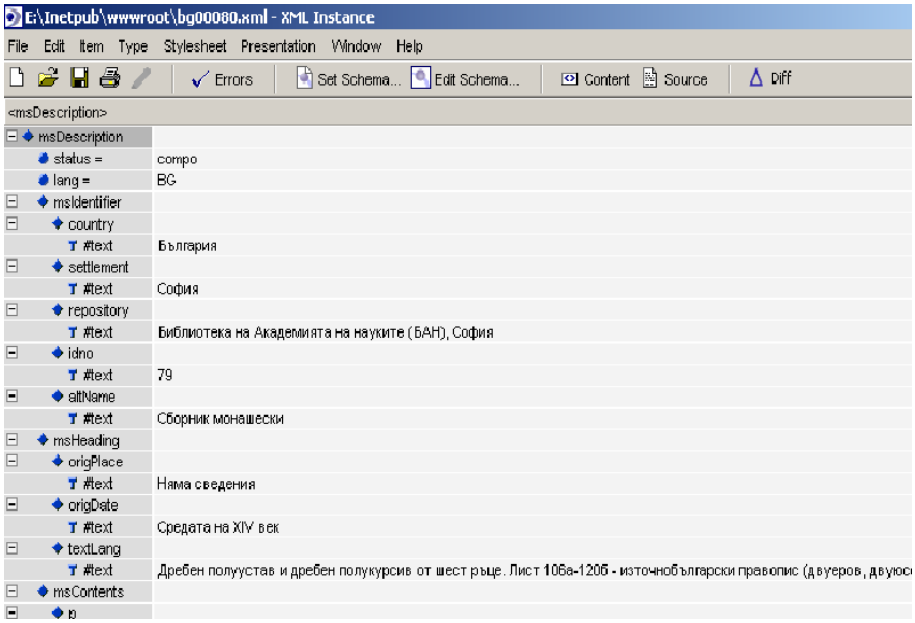

Fig. 5. An Example of Description Prepared in XEditMan Visualised in TurboXML editor

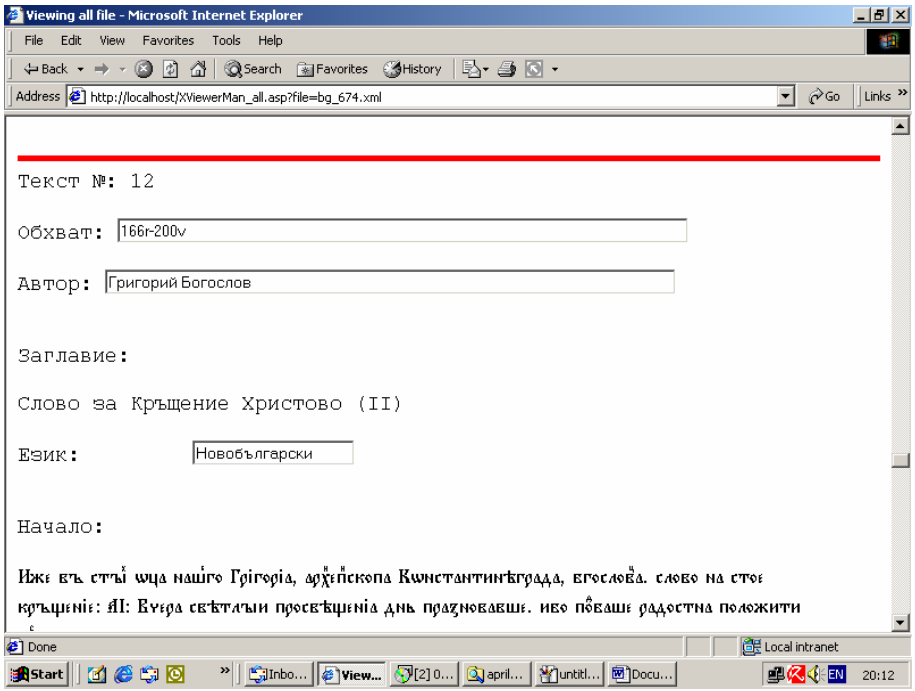

Fig. 6. An example of Visualized Data

Conclusion

The paper presented a brief overview of XML and the current trends in developing tools for its use. It formulated several basic requirements for the development of a specialized editor on mediaeval manuscripts, which guarantee faster and more accurate data entry.

It also presented the experience of the author in designing XEditMan, a specialized editor for manuscript descriptions. XEditMan was tested in the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences and is now used to enter data on Bulgarian manuscripts stored in Bulgaria. Two hundred descriptions are already available by the date of preparation of the paper (25 April 2004) based on the catalogue [Ikonomova et al., 1982].

This work is made as part of the current project Knowledge Transfer for the Digitization of Cultural and Scientific Heritage to Bulgaria, coordinated by the Institute of Mathematics and Informatics and supported by the Framework Programme 6 of the European Commission.

The basic idea is to provide in the next months a set of 800 manuscript descriptions which form about 1/10 of the manuscripts stored in Bulgaria. The first group of manuscripts, which was chosen, consists of Bulgarian manuscripts.

This work is extensible in two ways – more manuscripts could be added to the collection, and more data could be supplied at a later stage. For this reason, we believe that this initiative will contribute to the more adequate presentation of the cultural heritage of Bulgaria.

Bibliography

[Dobreva, 2000] 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 Perspectives, Sofia, Academic Publishing House, 2000, pp.. 91-98.

[Dobreva, Jordanova, 2000] M. Dobreva, M. Jordanova, *Some Psychological Aspects of Computer Modeling of Complex Objects,* In: A. Miltenova, D. Birnbaum (eds.), Medieval Slavic Manuscripts and SGML, Problems and Perspectives. Prof. M. Drinov Academic Publishing House, Sofia, 2000, pp. 295–310.

[Ikonomova et al., 1982] A. Ikonomova, D. Karadzhova, B. Christova, Bulgarian Manuscripts from 11 to 18 cent., stores in Bulgaria. Vol. 1. Sofia, 1982. (In Bulgarian – Икономова, А., Д. Караджова, Б. Христова. Български ръкописи от XI до XVIII век, запазени в България. Своден каталог, том I, НБКМ, София, 1982.)

[ISO, 1986] International Organization for Standardization, *ISO 8879: Information processing – Text and office systems - Standard Generalized Markup Language (SGML),* Geneva, ISO, 1986.

[Lund Principles, 2001] http://www.cordis.lu/ist/ka3/digicult/lund_principles.htm—*eEurope: creating cooperation for digitisation (Lund Principles)*

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

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

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

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

Author Information

Pavel Pavlov – Sofia University, Faculty of Mathematics and Informatics, Assistant Professor; 5 J. Bouchier Blvd, Sofia, Bulgaria; e-mail: pavlovp@fmi.uni-sofia.bg