Pavel I. Pavlov
(Faculty of Mathematics and Computer Science
Sofia, Bulgaria)

TOOLS FOR SEARCH IN ELECTRONIC COLLECTIONS OF MANUSCRIPT DESCRIPTIONS REALIZED IN THE INTEGRATED ENVIRONMENT XEditMan

Abstract: The paper presents an extension of the specialized editor XEditMan *(XML Editor for Manuscript Data)* which is an XML-oriented tool for editing and browsing catalogue descriptions of mediaeval manuscripts. The original version of XEditMan offers a friendly interface for entering data on mediaeval manuscripts, visualization and execution of queries to the descriptions that are already available. The descriptions are compatible with the document type definition (DTD) structure suggested by the project MASTER (*Manuscript Access through Standards for Electronic Records*) and adopted by the Text Encoding Initiative.

Recently XEditMan has been extended by tools for search of key words and phrases in a set of elements and attributes of all XML documents in an existing collection of manuscript descriptions prepared in accordance with the MASTER standard. The chosen set of elements and attributes includes: altName, watermarks, collation, foliation, scribe, script, medium, handDesc, origPlace, origin, acquisition, textLang, decoration, binding, condition, msContents, author, title, incipit, explicit, additions.

After the search tool is started, the user is asked to type the key word or phrase representing the query. Then he/she can choose a specific subset of the elements and attributes listed above. As a result the chosen XML elements and attributes of all documents in the collection containing words or phrases equal to the one given by the user should be properly visualized.

Key words: Manuscript Digitization, XML

1. Introduction

On-line access to 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, 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 are still not accessible.

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 specialized tools for data entry, processing and visualization. A set of such tools was developed by the author of this paper [1,2]. It helps to produce descriptions in electronic form faster and with better quality, and is of great help for the study of the material by target audiences with various interests. The latter is important in the light of personalization in the work with Internet.

This paper presents an extension of the specialized editor XEditMan which is an XML-oriented tool for editing and browsing catalogue descriptions of mediaeval manuscripts. The original version of XEditMan offers a convenient interface for entering data on mediaeval manuscripts, visualization and execution of queries to the descriptions that are already available. The descriptions are compatible with the document type definition (DTD) structure suggested by the project MASTER [3] and adopted by the Text Encoding Initiative. During the data entry the elements which are filled in appear in a sequence which is adopted in the manuscript cataloguing practice. The interface is in Bulgarian and this facilitates preparing electronic descriptions by people who are not acquainted in details with the DTD structure. The tool can be used also for visualization of single descriptions in two modes: complete descriptions or user-selected group of elements. Comparative study of multiple descriptions is achieved through database queries. XEditMan was used in the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences for the first mass data entry on Old Bulgarian manuscripts. Currently, a collection of 807 descriptions is available.

2. Global Search in the Entire Collection of Manuscript Descriptions

The visualization of single descriptions is convenient for users that are interested in a particular manuscript. Visualization may be alternated by editing of an existing description and vice versa. The search of a word or phrase in such description is limited by the facilities of the particular browser. In the course of use of XEditMan we realized that it is necessary to have a proper tool enabling the search of a word or phrase in the entire collection of manuscript descriptions. Such global search can give some global information about the contents of the manuscripts and their descriptions. An important question here is how to choose the set of most important elements and attributes in order to restrict the search activities in them due to effectiveness considerations. Interesting phrases as search queries should be the names of historic personalities, authors and scribes of manuscripts (although the names of authors usually are unknown), features of the manuscripts decoration, etc.

Mediaeval Bulgarian manuscripts stored in Bulgarian repositories are mostly with religious content. For that reason the additional texts (such as marginalia, scribblings, etc.) written by the scribes or owners of manuscripts are of great importance. They contain different kinds of information, e.g. records concerning historical events, social (mainly economic) or natural phenomena, events of life etc. Since there was no Bulgarian state at that time, it is impossible to find any official information of like nature. The records about natural phenomena might be concerned as a prototype of the information systems of contemporary hydrometeorological services while the records regarding solar eclipses and comets illustrate the features of pre-instrument astronomy. Records of economic nature contain information about crises caused by natural calamities or wars. They concern insufficiency of essential commodities and significant rise in their prices. There are lots of historical records with information about various wars, battles, risings, religious clashes etc. Varied data about visits of secular or cleric authority persons to cloisters and churches can be found as well. Fires, robberies, pestilences, long droughts and precipitations, floods, temperature irregularities etc. are also mentioned in the discussed additional texts. Some of them contain information about the persons commissioned the creation of the corresponding manuscripts.

Thus the additions made by the scribes in the margins of manuscripts can serve as an important source of information. Usually they contain records about natural and social phenomena getting people excited at that time. All these considerations predetermined the further extensions of XEditMan.

3. Tools for Search in a Set of Elements and Attributes

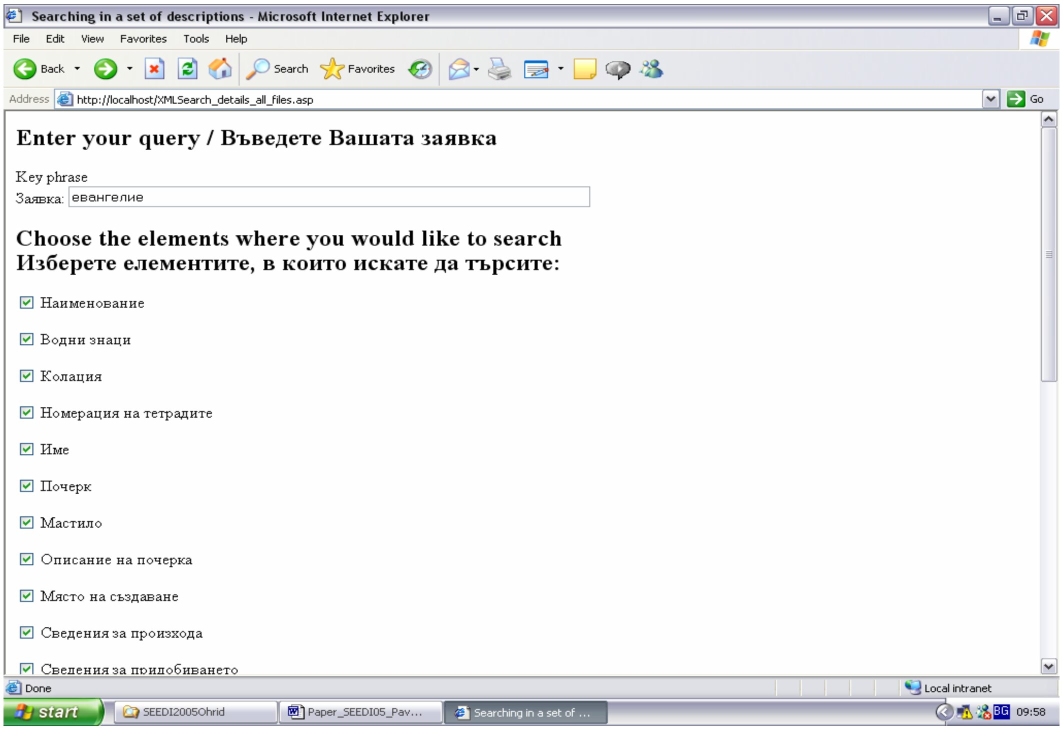
Recently XEditMan has been extended by tools for search of key words and phrases in a set of elements and attributes of all XML documents in an existing collection of manuscript descriptions prepared in accordance with the MASTER standard. The chosen set of elements and attributes includes: altName, watermarks, collation, foliation, scribe, script, medium, handDesc, origPlace, origin, acquisition, textLang, decoration, binding, condition, msContents, author, title, incipit, explicit, additions. We think that these elements and attributes are most informative.

Let us give some details about the features of these elements and attributes [3]:

- *altName* contains any form of alternative identifier used for a manuscript, such as a former catalogue number, ‘ocellus nominum’, or nickname;
- *watermarks* contains a detailed description of the watermarks identified in the paper of which a manuscript is composed;
- *collation* contains a description of how the leaves or bifolia are physically arranged;
- *foliation* describes the numbering system or systems used to count the leaves or pages in a codex;
- *scribe* gives a standard name or other identifier for the scribe believed to be responsible for this hand;
- *script* characterizes the particular script or writing style used by this hand;
- *medium* describes the tint or type of ink, e.g. ‘brown’, or other writing medium, e.g. ‘pencil’;
- *handDesc* gives a standard name or other identifier for the scribe believed to be responsible for this hand;
- *origPlace* contains any form of place name, used to identify the place of origin for a manuscript or manuscript part;
- *origin* contains any descriptive or other information concerning the origin of a manuscript or manuscript part;
- *acquisition* contains any descriptive or other information concerning the process by which a manuscript or manuscript part entered the holding institution;
- *textLang* describes the languages and writing systems used by a manuscript;
- *decoration* contains a description of the decoration of a manuscript;
- *binding* contains a description of one binding, i.e. type of covering, boards, etc. applied to a manuscript;
- *condition* contains a description of the physical condition of the manuscript;

- *msContents* describes the intellectual content of a manuscript or manuscript part either as a series of paragraphs or as a series of structured manuscript items;
- *author* identifies the primary author of the work or works contained in a manuscript;
- *title* contains the title of a work, whether article, book, journal, or series, including any alternative titles or subtitles;
- *incipit* contains the text of any *incipit* attached to a particular manuscript item, that is the opening words of a text, frequently used as a form of identifier for it; it may be preceded by one or more rubrics, and may be *defective*;
- *explicit* contains the text of any *explicit* attached to a particular manuscript item, that is, the closing words of a text or a section of a text, sometimes used as a kind of title, possibly followed by one or more rubrics or colophons;
- *additions* contains a description of any significant additions found within a manuscript, such as marginalia or other annotations.

After the search tool is started, the user is asked to type the key word or phrase representing the query (Fig. 1). Then he/she can choose a specific subset of the elements and attributes listed above (Fig. 1, Fig. 2). For search purposes we use a built-in function which performs textual comparison. We consider this option as most appropriate because it does not distinguish between capital and small letters and thus allows one to find more appearances of the searched text.


Searching in a set of descriptions - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/XMLSearch_details_all_files.asp Go

Enter your query / Въведете Вашата заявка

Key phrase
Заявка: евангелие

Choose the elements where you would like to search
Изберете елементите, в които искате да търсите:

☑ Наименование

☑ Водни знаци

☑ Колация

☑ Номерация на тетрадите

☑ Име

☑ Почерк

☑ Мастило

☑ Описание на почерка

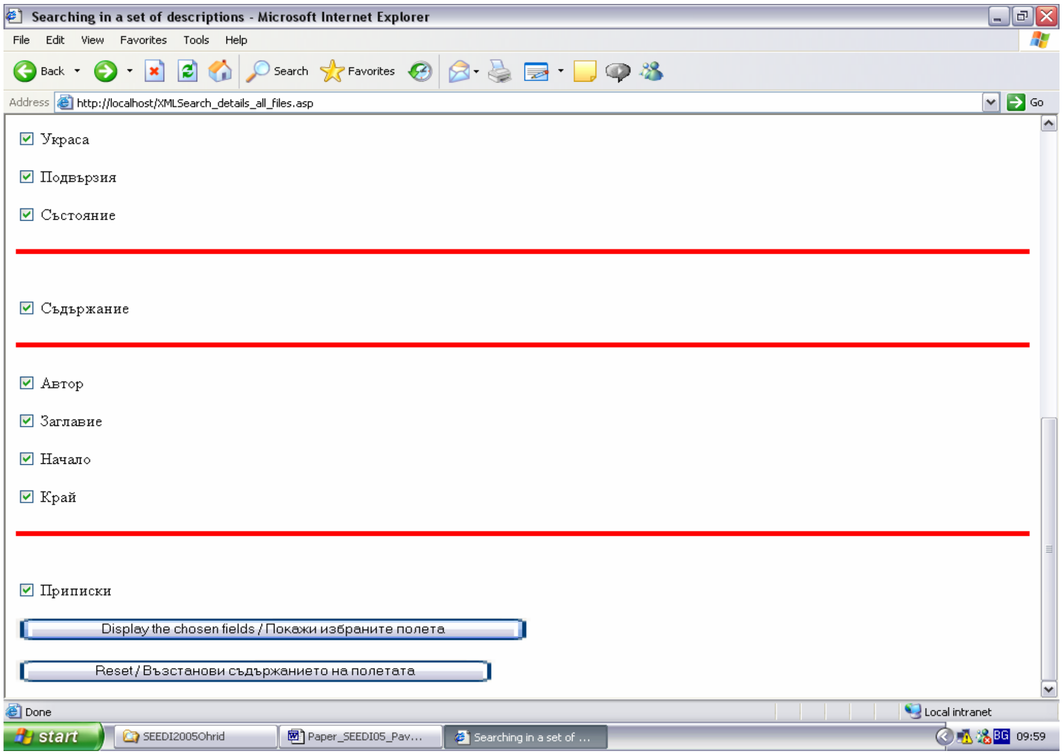
☑ Място на създаване

☑ Сведения за произхода

☑ Сведения за придобиването

Done
Local intranet
start
SEEDI2005Ohrid
Paper_SEEDI05_Pav...
Searching in a set of ...
BG
09:58

Fig. 1. An example user query


Searching in a set of descriptions - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://localhost/XMLSearch_details_all_files.asp
Go
Display the chosen fields / Покажи избраните полета
Reset / Възстанови съдържанието на полетата
Done
Local intranet
start
SEEDI2005Ohrid
Paper_SEEDI05_Pav...
Searching in a set of ...
BG
09:59

☑ Украса

☑ Подвързия

☑ Състояние

☑ Съдържание

☑ Автор

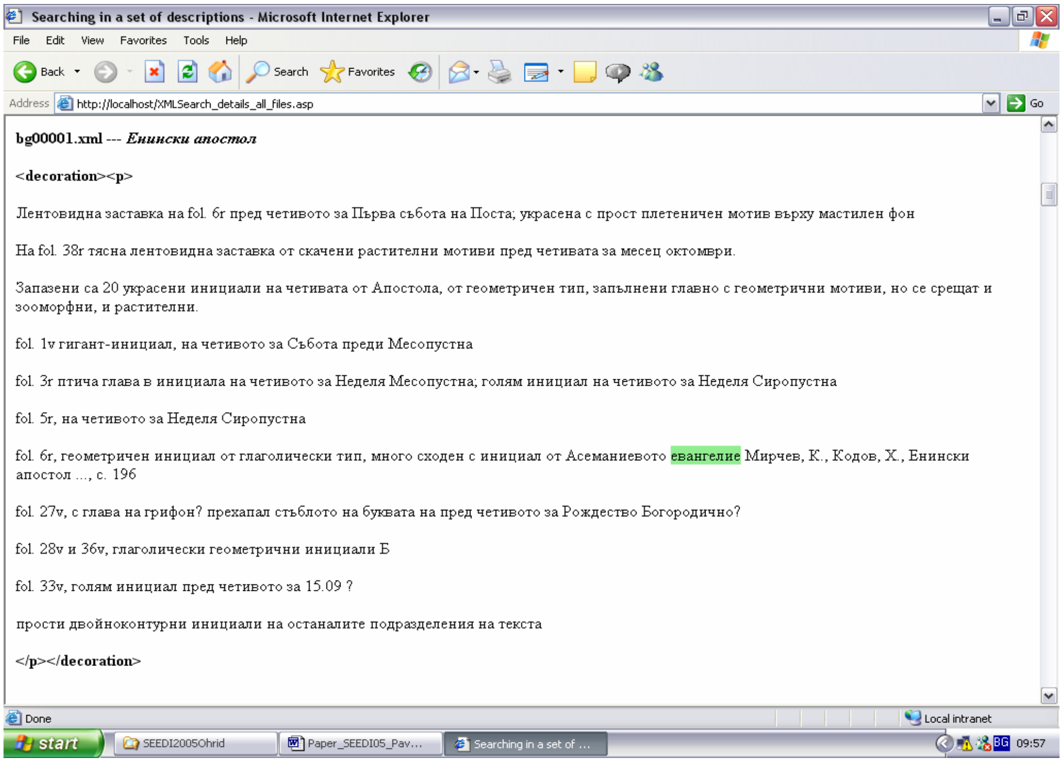
☑ Заглавие

☑ Начало

☑ Край

☑ Приписки

Fig. 2. An example user query (continuation)


Searching in a set of descriptions - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://localhost/XMLSearch_details_all_files.asp
Go
Done
Local intranet
start
SEEDI2005Ohrid
Paper_SEEDI05_Pav...
Searching in a set of ...
BG
09:57

bg00001.xml --- ***Енински апостол***

<decoration><p>

Лентовидна заставка на fol. 6r пред четивото за Първа събота на Поста; украсена с прост плетеничен мотив върху мастилен фон

На fol. 38r тясна лентовидна заставка от скачени растителни мотиви пред четивата за месец октомври.

Запазени са 20 украсени инициали на четивата от Апостола, от геометричен тип, запълнени главно с геометрични мотиви, но се срещат и зооморфни, и растителни.

fol. 1v гигант-инициал, на четивото за Събота преди Месопустна

fol. 3r птича глава в инициала на четивото за Неделя Месопустна; голям инициал на четивото за Неделя Сиропустна

fol. 5r, на четивото за Неделя Сиропустна

fol. 6r, геометричен инициал от глаголически тип, много сходен с инициал от Асеманиевото евангелие Мирчев, К., Кодов, Х., Енински апостол ..., с. 196

fol. 27v, с глава на грифон? прехапал стъблото на буквата на пред четивото за Рождество Богородично?

fol. 28v и 36v, глаголически геометрични инициали Б

fol. 33v, голям инициал пред четивото за 15.09 ?

прости двойноконтурни инициали на останалите подразделения на текста

</p></decoration>

Fig. 3. Some search results

4. Conclusion

Our first experiments have been carried out with the mentioned above collection of 807 catalogue descriptions of mediaeval Bulgarian manuscripts stored in Bulgaria. The search in this collection is completed in approximately 5-6 seconds. We consider the results of these first experiments with the search tool of XEditMan as promising.

Acknowledgements. This work has been funded by the EC FP6 Project “Knowledge Transfer for Digitisation of Cultural and Scientific Heritage in Bulgaria” (KT-DigiCULT-BG).

References

[1] M. Dobreva, P. Pavlov, *How to Enter Data on Mediæval Bulgarian Manuscripts More Easily, or XEditMan: An XML Editor for Manuscript descriptions*, Scripta & e-Scripta 2 (2004), Sofia, “Boyan Penev” Publishing Center, pp. 79–95.

[2] P. Pavlov, *XEDITMAN: A XML Editor for Manuscript Descriptions and its Implementation for Cataloguing of Bulgarian Manuscripts,* Review of the National Centre for Digitisation 5 (2004), pp. 49–58.

[3] L. Burnard, *Reference Manual for the MASTER Document Type Definition*, http://www.tei-c.org.uk\Master\Reference\index.html.

pavlovp@fmi.uni-sofia.bg