An Approach to Representation of Digitized Archival Collections

Àâòîð Ìèëåíà Ïåòðîâà Äîáðåâà
28.06.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 15.07.2008 ã.

Òåçèñû â ôîðìàòå RTF (79.58 kB 2008-07-15 16:31:54) Òåçèñû â ôîðìàòå PDF (98.77 kB 2008-07-15 16:32:26)

1. Introduction

The role of information technologies in the digital preservation of collections of handwritten, typewritten and printed archival documents has rapidly grown in the recent years. Special attention is being paid to enhanced access to digitized collections of documents.

In particular, the number of electronic publications of archival collections which are of interest to narrow domain specialists (archivists, historians, linguists etc.) and to the general citizen is increasing.[1] However, these electronic publications typically provide access tools oriented to the “traditional” archivist’s point of view: it is only possible to browse the full archival structure traditional for the host country, so searching for documents is very difficult and the given search means are too limited.

This paper presents an ongoing project which aims to develop a methodology and software tools for providing semantics-oriented, web-based access to distributed digitized archival collections. These collections are heterogeneous, i.e., they may include diverse types of materials (official handwritten, typewritten or printed documents, letters, photographs, newspapers, maps, etc.) and the texts of the documents within them may be written in different languages. The experiments have been performed on a collection of archival documents from the period of the establishment of the Sofia Municipal Government (1878—1879).

2. Representation of the Archival Documents

In this paper we present an ongoing effort aimed at creating an electronic version of an archival collection which consists of approximately 980 original handwritten documents from the period during and after the Russo-Turkish war (1877—1878). The documents within the collection are of great scientific, historical, and social value and are of interest to archivists, historians, linguists, etc. For these reasons it is essential to include in the electronic version of our collection not only digital images of the archival documents but also structured electronic transcriptions of their full texts and proper descriptions of the collection as a whole as well as descriptions of its parts (known as archival units) and all particular documents in it.

2.1. Description of the Structural Parts of the Archival Collection

The structure of Bulgarian archival record has four levels of hierarchy: archival funds, inventory lists, archival units and individual documents. The descriptions at all levels have been structured and accompanied with proper sets of metadata according to the requirements of the Encoded Archival Description (EAD) encoding scheme.[2]

2.2. Representation of Electronic Transcriptions of Full Texts of Archival Documents

We decided to store two different digital objects corresponding to each original archive document: its digital image in PDF and an electronic transcription of its full text in XML format. The digital images of the original documents are intended mainly for visualization while the electronic transcripts of the documents and their EAD encoded descriptions will be used for document retrieval. For the representation of the structured electronic transcriptions of the full texts of archival documents we use the Text Encoding Initiative (TEI).[3] We explored the structure and the contents of various kinds of documents within the collection (instructions, orders, reports, records of sessions, letters, requests, petitions etc.) to create a generalized model of these documents. A proper set of elements and attributes from the TEI document type definition was adopted to describe this model.

3. Access to the Collection

The outcome of the project will give the user the opportunity to switch between two types of interface to the chosen collection. The first one is based on the principles of the “typical” archivist’s view to an archival collection. The second type of provided on-line access to the collection may be described as the semantics-oriented one.

The interface to the archival collection oriented to the standard archivist’s point of view allows the user to browse the hierarchical structure of the collection as a whole. At the archival fond and inventory list levels the user has an access to the EAD-encoded description of the corresponding unit (in XML format) and to a properly visualized form of the same metadata in PDF.

The user interface at archival unit level allows one to browse five different forms of each particular document in the corresponding archival unit: the EAD encoded description of the document (in XML format), a proper visualization of this description in PDF, the TEI encoded electronic transcription of the full text of the document (in XML format), a proper visualization of the electronic transcription of the document in PDF, and a digital image of the original document (again in PDF). Short historical data accompany this type of interface to the collection.

The other type of provided access to the discussed archival collection is based on the use of explicitly represented knowledge describing different aspects of the semantics of the collection as a whole and its structural parts. A set of access tools supporting various types of search and retrieval (chronological, oriented to the kinds of documents within the collection, subject oriented etc.) are under development. The search engines of most of these tools use the values of the corresponding elements of the TEI encoded versions of archival documents. In particular, subject-oriented document retrieval is based on the use of the semantic annotation of the documents.

The development of ontologies is still a difficult task because so far there are no common platforms and verified methods which would prescribe what procedures should be followed in the

process of creating an ontology. In the discussed project we use a subject ontology (covering the main types of municipal activities) especially developed for the purpose and prepared using Protégé in OWL.[4]

4. Conclusion

This paper presents a work in progress directed to the exploration of some open questions concerning the development of proper mechanisms and tools providing adequate web-based access to digitized archival collections. The most valuable expected results of our project could be formulated as follows:

.
A methodology for application of international standards, ontological knowledge, and Semantic web technologies for the development of software tools providing semantics-oriented access to heterogeneous multilingual collections of archival documents;

.
A model and a prototype of a website which gives the users an interface supporting various types of access to a chosen archival collection.

Acknowledgements

This work has been funded by the EC FP6 Project KT-DigiCULT-BG coordinated by IMI-BAS. The authors are grateful to Dr. M. Driscoll from Copenhagen University, Denmark, for his useful advices concerning the TEI encoding of the electronic copies of archival documents.

Ïîäõîä ê ïðåäñòàâëåíèþ îòêðûòûõ àðõèâíûõ êîëëåêöèé

Ì. Ï. Äîáðåâà1,3, Ï. È. Ïàâëîâ2, Ì. Ì. Íèøåâà-Ïàâëîâà2,3

1Óíèâåðñèòåò
Ñòðàòêëàéä, Ãëàçãî, Øîòëàíäèÿ, Âåëèêîáðèòàíèÿ

2Ôàêóëüòåò
ìàòåìàòèêè è èíôîðìàòèêè, Ñîôèéñêèé óíèâåðñèòåò, Ñîôèÿ, Áîëãàðèÿ

3Èíñòèòóò
ìàòåìàòèêè è èíôîðìàòèêè Áîëãàðñêîé ÀÍ, Ñîôèÿ, Áîëãàðèÿ

Äîêëàä ïðåäñòàâëÿåò ïðîåêò,
îñíîâíîé öåëüþ êîòîðîãî ÿâëÿåòñÿ ðàçâèòèå ìåòîäîëîãèè è ïðîãðàììíûõ ñðåäñòâ ñåìàíòè÷åñêè-îðèåíòèðîâàííîãî âåá-äîñòóïà ê äîêóìåíòàì îöèôðîâàííûõ àðõèâíûõ êîëëåêöèé. Äëÿ èñïëåäîâàíèÿ è ýêñïåðèìåíòîâ áûëè èñïîëüçîâàíû äîêóìåíòû ôîíäà

Âðåìåííîãî ðóññêîãî ïðàâëåíèÿ ïîñëå îñâîáîæäåíèÿ Áîëãàðèè 1878-1879 ãã., êîòîðûå õðàíÿòñÿ â Ãîñóäàðñòâåííîì àðõèâå Áîëãàðèè.

[1] See
for example Brown Archival & Manuscript Collections Online (BAMCO) site. URL: http://dl.lib.brown.edu/bamco/.

[2] The Encoded Archival Description (EAD). URL: http://www.loc.gov/ead/.

[3] The Text Encoding Initiative
(TEI). URL: http://www.tei-c.org/.

[4] OWL Web Ontology Language Overview. W3C
Recommendation, 10 February 2004. URL: http://www.w3.org/TR/owl-features/,
last accessed on April 4, 2007.