Multimedia database of the cultural heritage of the Balkans

Ivana Tanasijević[1], Biljana Sikimić[2], Gordana Pavlović-Lažetić[1]

[1]Faculty of Mathematics, University of Belgrade,

[2]Institute for Balkan Studies of Serbian Academy of Sciences and Arts,

11000 Belgrade, Serbia

E-mail: {ivana, gordana}@matf.bg.ac.rs, Biljana.Sikimic@bi.sanu.ac.rs

Abstract

This paper presents a system that is designed to make possible the organization and search within the collected digitized material of intangible cultural heritage. The motivation for building the system was a vast quantity of multimedia documents collected by a team from the Institute for Balkan Studies in Belgrade. The main topic of their research were linguistic properties of speeches that are used in various places in the Balkans by different groups of people. This paper deals with a prototype system that enables the annotation of the collected material and its organization into a native XML database through a graphical interface. The system enables the search of the database and the presentation of digitized multimedia documents and spatial as well as non-spatial information of the queried data. The multimedia content can be read, listened to or watched while spatial properties are presented on the graphics that consists of geographic regions in the Balkans. The system also enables spatial queries by consulting the graph of geographic regions.

Keywords: cultural heritage, multimedia annotation, native XML database

1. Introduction

Preserving the national heritage is of great importance for understanding specifics of a nationality. With the significant increase of interest in the online usage of information, there is a strong requirement for making it available to the public. A large number of documents were created and still exist in the non-digitized form. Therefore, there is a growing need for digitizing and organizing those collections of materials. Digitalization is the most appropriate solution for several reasons (Digital Libraries, i2010):

- Digital documents are easier to share and hence are more accessible to citizens or researchers who could increase the optimality of the usage of collected information
- A digital form is of significant importance for materials that can be damaged over time or lost for any reason, so it enables the preservation of valuable material that could be used for other purposes and accessed by future generations.

Compiling a collection of documents that represents the national heritage is a long, expensive and demanding task. Digitization of such a material is performed by scanning the documents and images and re-typing the text. It also includes various types of word processing or processing the audio and video materials. Another issue that must be taken into account is that it is often performed by making a copy of someone else's property. Therefore, legal issues can also present a problem that requires special attention.

Large increase in digitization of national and scientific heritage has been noticed on the territory of Europe, including Serbia, in the last few years. There are annual conferences on digitization of national heritage held at the Faculty of Mathematics (NCD)[1] , Belgrade, SEEDI[2] conferences and yet another effort in the same direction was a conference dedicated to preserving historical, cultural and scientific heritage held by the Faculty of

[1]http://www.ncd.matf.bg.ac.rs

[2]http://seedi.ncd.org.rs

Philology in Belgrade[3] .

This paper presents a system that is designed to organize a large amount of digitized multimedia material of intangible cultural heritage. The main motivation is the collection of documents that have been collected by a team from Balkan Studies in Belgrade for last twelve years. This system enables annotation and organization of this large collection into native XML database. The data can be annotated by spatial as well as non-spatial information, thus in this system both type of the informations of the queried data can be presented. Assigned non-spatial data are presented in text form, while spatial data are presented on a graphic that consists of geographic regions in the Balkans. In addition to this informations, multimedia content of those files can be read, listened to or watched. The system also enables spatial queries by consulting the graphic of geographic regions. The system is implemented in Java.

2. Related work

Multimedia Content Management System (MILOS)[4] is developed for a maintenance of heterogeneous documents and their associated metadata, which allows storing multimedia documents and efficient search. It is a general purpose software component tailored to support design and effective implementation of digital library applications. MILOS supports the storage and content based retrieval of any multimedia documents whose descriptions are provided by using arbitrary metadata models represented in XML. It is based on advanced techniques with XQuery language and implements features for image similarity search, text search, categorization and meny more.

Another management system for systematization of texts and images and possibly linking between different systems, which allows the contrast to the systems considerably more versatile information retrieval is Museum24[5] . The system is easily adaptable to any cultural heritage material and a museum web publishing. It uses tools like XML, RDF, OWL and MPEG-7. Projects like MultimediaN[6] and eChase[7] also explore the use of semantics enrichment in a cultural heritage domain.

The European Library[8] is a free service that provides a vast collection of materials from many disciplines. It contains library material, books, maps, photographs, audio and video material. It offers access to the resources of the 48 national libraries of Europe in 35 languages.

3. Multimedia collection of documents

The collected material presented in this paper includes interviews in different languages and dialects used in various places in the Balkans. Some of them were recorded in Hungary (Serbian and Bunjevci vernaculars), then in Croatia, Bosnia and Herzegovina, and Bulgaria (Romanian speaking Roma). A large corpus of interviews with Bunjevci, Croats, different types of Roma which speak Albanian, Romani or Romanian, Romanians, Vlachs, Czechs, Bulgarians, was collected in the territory of Serbia. It also includes the entire territory of Kosovo and Sandzak where also Muslims were interviewed, as well as speakers whose first language is Serbian.

Team of experts from the Institute for Balkan Studies of Serbian Academy of Sciences and Arts was involved in collecting the data. During the time, in this work were envolved about fifteen experts. The material consists of a large number of photographs, audio and video recordings. It is mostly in the form of conversations and it comprises approximately 2000 hours of oral date. In addition to these types of materials there are transcripts of the recorded interviews. Some of them have been translated into several languages. Each audio and video material has an introduction (in the form of a protocol) in the written form which describes the content of the material. Some materials are associated to the already published papers which result from the research that was made over them. Some examples of the materials are shown in Figure 1 and Figure 2.

[3] http://digitalheritage.fil.bg.ac.rs/?lan=en
[4] http://milos.isti.cnr.it
[5] http://www.museo24.fi
[6] http://e-culture.multimedian.nl
[7] http://www.echase.org
[8] http://www.theeuropeanlibrary.org

Figure 1: National clothes and customs

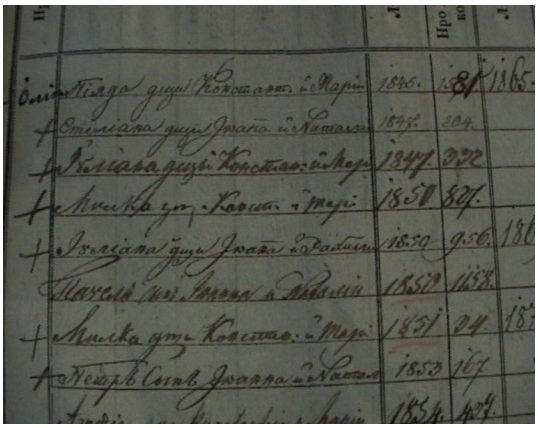


Figure 2: Digitized document of an old manuscript

Till this moment, a lot of studies, books, monographs and hundreds of papers were published about the collected material. Some papers can be found in (Sikimić, 2006), (Sikimić, 2007) and (Sikimić, 2009). The main motivation of the research presented is to help fully digitize the corpora, and then to systematize and organize the collection so that it becomes available to the wider academic community and searchable by keywords or by geographical markers in the Balkans.

The possible occurrence of similar properties in different geographic regions can lead to some conclusions about possible migration of certain groups of people. Further, it can show all the characteristics of the groups that are territorially close. We also plan to make the content of some material available to the general public.

4. Methodology

Different kinds of non-digital materials were first converted into the digitized form. The photographs have been scanned. The audio and video materials have been converted from tapes and films to digital files. Later, the material was created by digital cameras, so it was already in the digital form which makes further processing much simpler. Most of the introductions have been re-typed.

In order to organize and store the data, native XML database eXist[9] is used. The database consists of a collection of XML documents that organizes each of the stored material into structured data. The XML technologies are the most suitable for this kind of problems (Tošić, 2005). The corpora was initially organized through a system of folders and files without a unique and uniform organization that is required for the efficient search. To find the optimal way to organize them, it was necessary to analyze the type and content of multimedia materials and to find the key labels with which material should be annotated. Some analyses were performed in order to identify groups of features which can be used to mark the material in the best way. Those groups consist of place and region in which the material was recorded, nationality and religious affiliation of interviewees, languages in which the conversation was recorded, topics discussed and authors who participated in the surveys

5. Architecture of the system

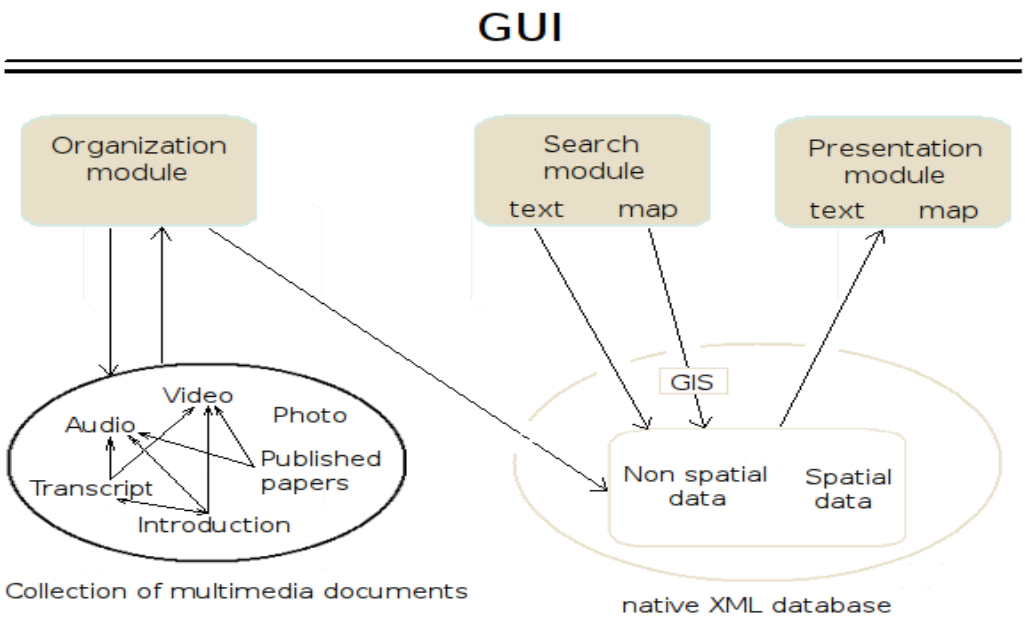
The system provides three modules: for annotation of data and updating the database, for searching the database and for presenting the content that is relevant to the searching criteria. The organization of the system is shown in Figure 3.

Figure 4 shows the panel for storing new data. A simple interface designed for authors to select the documents, annotate it and store materials in the existing database is created. This feature greatly facilitates the classification and systematization of such a large corpus of the multimedia documents.

The system also enables the modification of the materials or their new annotation.

The annotation tags that are associated with the material are structured in such XML records and then stored into the database using XUpdate queries (Xquery, 2005).

[9] http://exist-db.org



.

Figure 3: Architecture of the system

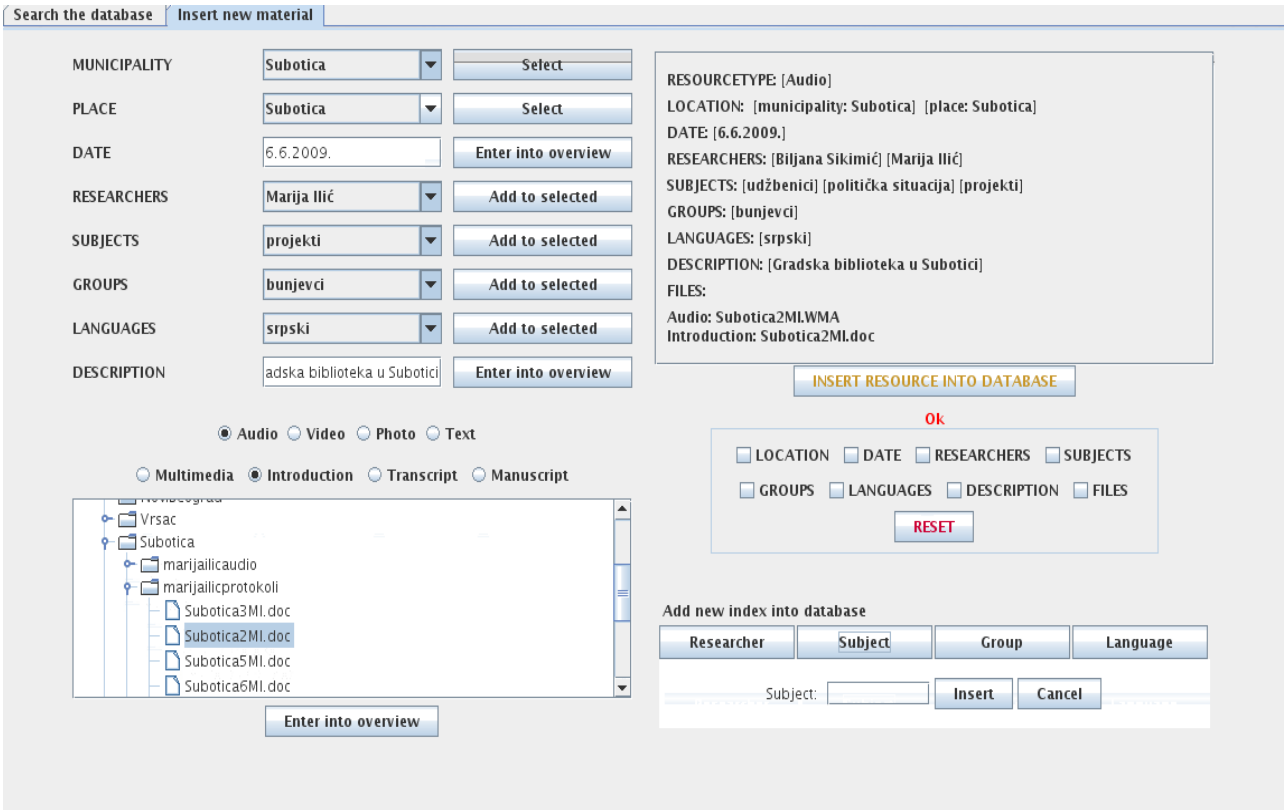


Figure4: The annotation and data entry

In addition to the non-spatial data that has been mentioned above, there are spatial data such as coordinates of places in the Balkans and municipalities in Serbia. These coordinates are also stored in a structured form in the database in Geography Markup Language (GML) files. GML is a language that primarily describes the geographic entities and sets of such entities

6. Searching and presenting data

Data that are already annotated can be searched by annotation criteria. There are two types of searches:

- by non-spatial attributes, such as language, nationality, topic, author
- by location where a material was created

Non-spatial criteria can be selected using the graphical user interface. An example of a search is shown in the Figure 5. XQuery queries, that are used in searching the database, are well structured queries and the efficiency and optimization of the search was strongly taken into account. Graphical user interface for searching consists of six classes by which a search can be performed. Every single class can be part of a search requirements or can be excluded from a query. Those classes are: *type of file*, *municipality*, *place*, *author*, *language*, *group* and *subject*, where type of file can be “text”, “audio”, “video” and “photo”. There is a possibility to support other file types and it would require some kind of adaptation.

When the search is completed, the results, which consist of all information about the relevant materials in addition to the paths to these materials, are singled out and presented in the graphical and non-graphical way. As the corpora consists of multimedia materials, each of the document can be viewed or listened, as well as transcripts and introductions associated with them. Conversations that are recorded were held in the different dialects, which are often incomprehensible to people who do not speak that dialect, but speak the same language. Thus, some transcripts are translated into several languages, so the system provides a parallel view of transcripts in two or more languages. The graphical representation of the Balkans' map that is second, spatial part of the queried data, indicates regions where the resulting materials were created. In this way a more complete picture of the presence of certain customs or languages held by certain groups of local communities from the classification of interest is made.

On the other hand, the interactive map of the Balkans also provides search in the opposite direction - from the map to non-spatial characteristics. Selecting a particular region or place on the map makes all materials related to that location available for presentation. By this way, search is performed by one of two criteria mentioned above, and those are *places* or *municipalities.* For spatial calculations eXist's spatial module is used, which consists of many methods that can perform various topological and metric calculations that are part of the Geographic Information System (GIS).

7. Implementation

This system is written in Java. It consists of client and server parts that communicate using Transmission Communication Protocol (TCP). The server part is responsible for the database queries and for communication with the client part. The client part is responsible for displaying graphical user interface and for communication with end client over World Wide Web's Hypertext Transfer Protocol (HTTP). Also, it communicates with the server part.

The server part is implemented as a process that waits for requests for TCP connections on certain port. When the next client connects, this process makes a new thread that deals with further requests from the client. On the other hand, the client part is implemented as an Java applet built into the html page. When a browser sends a request for html page via HTTP protocol, the client applet tries to connect by the TCP protocol to the server part of application.

The server part consists of three major working units: Server, ServerThread and ServerDB. The ServerDB implements methods which process and prepare queries that are requested by client. Each method is responsible for one type of query, where the parameters of query depend of the arguments of a given method. A thread that is responsible for a given client connects to the eXist native XML database using a Simple Object Access Protocol (SOAP). SOAP is a protocol for information exchange between same or different operating systems in a network communication by using HTTP protocol and its XML files. It specifies exactly how to encode a HTTP header and a XML file so that a program in one computer can call a program in another computer and pass the information.

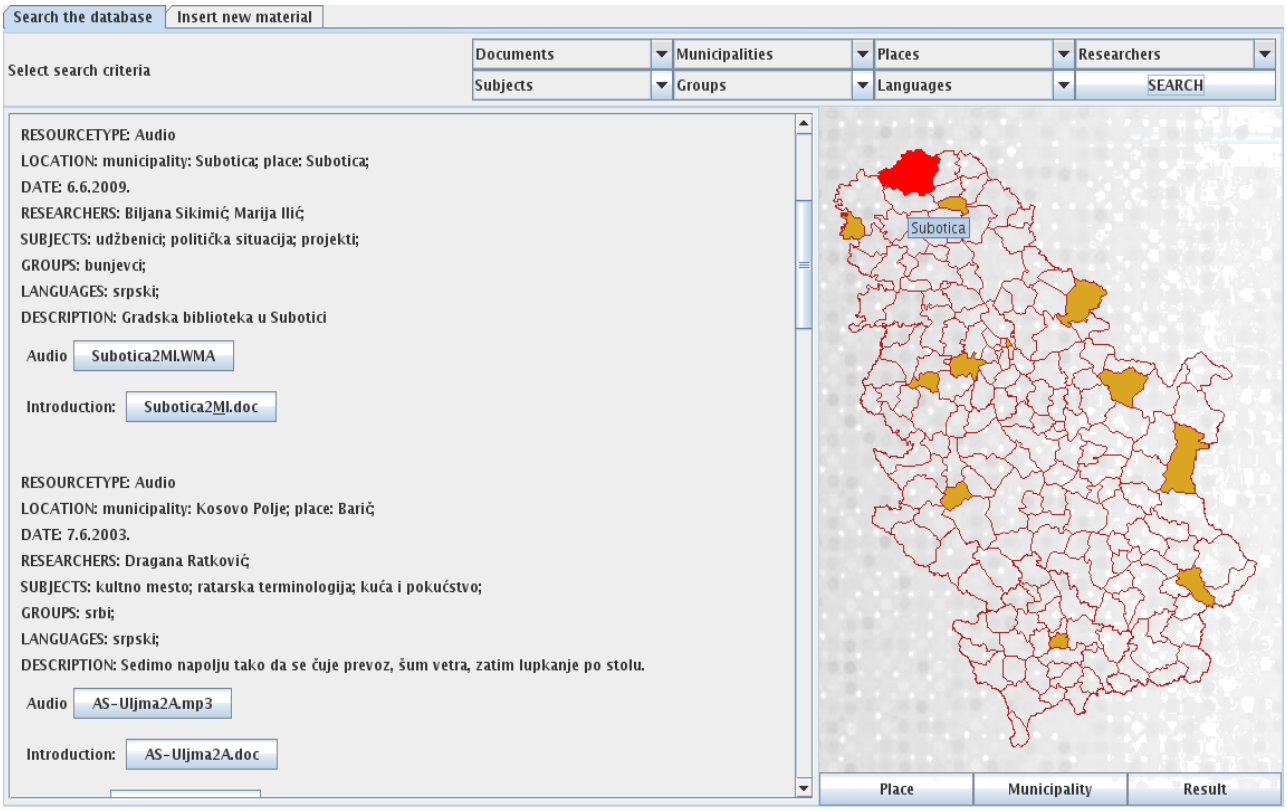


Figure 5: Example of search and displaying results

It also specifies how the called program can return a response. Server thread then executes one or more queries and returns the answer to the client. In order to use services for eXist database, certain jar files need to be included in a classpath of the server side of the application. Client applet consists of two major units, namely ClientGUI and ClientDB. The ClientGUI forms a complete graphical interface for presenting, searching and inserting data into a database. The ClientDB defines methods corresponding to methods from the ServerDB. This methods store all of the arguments in an object, then send the object to the server and wait for the query result. Some of the methods which perform queries are:

```
public Vector<String> getNames(Integer type);
public Integer insertLanguage(String language);
public Integer insertObject(ArrayList<Object> object);
```

where object consists of:

```
Integer type, Vector<String> location,  Vector<String> researchers,
Vector<String> languages, Vector<String> groups, Vector<String> subjects,
Vector<Vector<String>> files
```

The database itself consists of three parts:

- an index file
- a file with a spatial data and coordinates of municipalities
- the information about annotated materials or those that are entered into the database

An example of a record that describes a place in Serbia has a next form:

```
<place>
    <id>184</id>
    <name>Blace</name>
    <coordinates>
        <gml:Point>
            <gml:coordinates>
                21.28750560,43.29997589
            </gml:coordinates>
        </gml:Point>
    </coordinates>
</place>
```

Queries used are written in XQuery language. An example of a query that returns coordinates of Blace is:

```
for $m in doc("places.xml")//place
where $m/name/text()="Blace"
return $m//gml:coordinates/text()
```

A result of this query is a string 21.28750560,43.29997589, which server then parses and transmits to the client.

The collection of documents consists of photographs, audio and video files, as well as associated introductions and transcripts. In addition, links to hundreds of published papers can be assigned to these files. These files are grouped by type and their counts are shown in Table 1.

File type	Count
Photo	9318 files
Audio	3135 files, average length is 45 minutes
Text	844 files
Video	362 files, average length is 20 minutes

Table 1: Structure of the collection of documents

Here is one example of a record that presents one audio file:

```
<audio>
  <file id="1">
    <location>
      <municipality id=”356”/>
      <place id=”17”/>
    </location>
    <date>11.06.2006.</date>
    <description>Traditions during the holidays.</description>
    <researchers>
      <researcher id=”1”/><researcher id=”5”/>
    </researchers>
  <groups><group id=”7”/><group id=”4”/></groups>
  <languages><language id=”1”/></languages>
   <subjects><subject id=”10”/><subject id=”4”/></subjects>
  <paths>
    <multimedia>/DABI/Apatin/bs.mp3</multimedia>
     <introduction>/DABI/Apatin/bs.1.doc</introduction>
    <transcript>/DABI/Apatin/bs.1.pdf</transcript>
     <manuscript>/DABI/scripts/apatin4.pdf</manuscript>
  </paths>
 </file>
</audio>
```

This record shows that the ids of the key tags from the index file are kept, rather than the names themselves.

At the moment, the database containds the files that are recorded in Kosovo in year 2003., Ibarski Kolasin, and most of the files with Bunjevci people. Whole material should be annotated through time. At this phase, the application can be found on the address http://science.matf.bg.ac.rs/dabi. It is accessible with account which can be obtained via email. If you are interested in using the application please contact the corresponding authors.

The source code of this application is not publicly available which may change at some point in the future.

8. Conclusion and future work

The most important characteristics of the system presented for handling multimedia documents of the cultural heritage of the Balkans is that it provides a searchable organization of a large corpus of multimedia documents, assocciating every material with a relevant region on the map. Thus enables a better point of view of the content, which opens up possibilities for further research and discovery of new knowledge. The system may be applied to other corpora of multimedia documents as well and can be also used for other initiatives like this. The application presents the certain region, but it can be easily modified to fit the data for some other region that need to be organized in a searchable form.

The future plan is to enable automatic annotation of the documents. As this process can not be fully automated, some kind of revision is needed. The basic idea is to search the content of the affiliated introductions and so provide possible relevant tags. Another issue that we plan for the future work is a search for some word or sequence of words through the text documents in the database. Multimedia documents which can be obtaind by this query will be ones that refer to those containing some word or have context that is relevant to the given questioned sequence.

The native XML database used is the most suitable for this type of problem and data and is capable for easy organizing and searching. XML files, which are basic units in which the data are stored, are easy to create, read and access, which means that they can be viewed from the outside of the database and easily understood in accordance with the document scheme.

9. References

Crofts, N., Doerr, M., (2003). Comprehensive Introduction to CIDOC CRM, ICOM/CIDOC DSG, web presentation, http://cidoc.ics.forth.gr/comprehensive_intro.html

Digital national library of Serbia, http://eng.digital.nb.rs/

Digital Libraries, i2010: http://ec.europa.eu/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf

The European Library, http://www.theeuropeanlibrary.org/

eChase, http://www.echase.org/

eXist, open source native XML database, http://exist-db.org/

MILOS , Multimedia Content Management *System*, *http*://milos.isti.cnr.it/

MultimediaN, http://e-culture.multimedian.nl/

NCD, National center for digitalization, http://www.ncd.matf.bg.ac.rs

Pavlović-Lažetić, G. (2007). Native XML databases vs. relational databases in dealing with XML documents, Kragujevac J.Math. 30, 181-199

Scientific conference Digitalisation of Cultural and Scientific Heritage, University Repositories and Distance Learning, Belgrade, Serbia, http://digitalheritage.fil.bg.ac.rs/?lan=en

SEEDI, South-Eastern European Initiative, http://seedi.ncd.org.rs/

Sikimić, B. (2006). *Krvna Žrtva, Transformacije jednog rituala*, Beograd

Sikimić, B., Hristov, P. (2006). *Курбан на Балкану*, Beograd

Sikimić, B. (2009). *Politika transkripcije i interperformativnost, Moć Književnosti*, in memoriam Ana Radin, Beograd

Sikimić, B., (2009). *Politika transkripcije i interperformativnost, Moć Književnosti*, in memoriam Ana Radin, Beograd

Szász, Cultural Heritage on the Semantic Web - the Museum24 project, http://www.museo24.fi

Tosić, D. (2005). *XML-tehnologies and digitalization, Review of the National Center for Digitization*, 1 -12, http://elib.mi.sanu.ac.rs/files/journals/ncd/3/d001download.pdf

XQuery 1.0 (2005). An XML query language, candidate recommendation, http://www.w3.org/TR/2005/CR-xquery-20051103/