A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, Miloš Utvić

A tool for enhanced search of multilingual digital libraries of e-journals

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, Miloš Utvić
University of Belgrade, Serbia
Studenski trg 1, 11000 Belgrade
E-mail: ranka@rgf.bg.ac.rs, cvetana@matf.bg.ac.rs, ivano@rgf.bg.ac.rs, aleksandra@unilib.bg.ac.rs, misko@matf.bg.ac.rs

Abstract

This paper outlines the main features of Bibliša, a tool that offers various possibilities of enhancing queries submitted to large collections of TMX documents generated from aligned parallel articles residing in multilingual digital libraries of e-journals. The queries initiated by a simple or multiword keyword, in Serbian or English, can be expanded by Bibliša, both semantically and morphologically, using different supporting monolingual and multilingual resources, such as wordnets and electronic dictionaries. The tool operates within a complex system composed of several modules including a web application, which makes it readily accessible on the web. Its functionality has been tested on a collection of 44 TMX documents generated from articles published bilingually by the journal INFOtecha, yielding encouraging results. Further enhancements of the tool are underway, with the aim of transforming it from a powerful full-text and metadata search tool, to a useful translator's aid, which could be of assistance both in reviewing terminology used in context and in refining the multilingual resources used within the system.

Keywords: multilingual digital libraries, query expansion, TMX

1. Motivation

In this paper we outline the main features of Bibliša (http://hlt.rgf.bg.ac.rs/Biblisha), a tool developed within the Human Language Technology group at the University of Belgrade, aimed at enhancement of search possibilities in multilingual digital libraries of e-journals. The tool offers cross-lingual search functions for large collections of aligned texts, which enable users to compose queries both with simple and multiword keywords in more than one language. In addition to that, user queries can be expanded, both semantically and morphologically, the latter being very important in highly inflective languages, such as Serbian.

Open access Serbian scientific journals are increasingly present on the web. However, they are in general monolingual, but even then full-text search is either not supported or remains very limited (e.g. string search). INFOtheca (http://infoteka.bg.ac.rs/), which covers the field of Library and Information Sciences, is one of the few journals that publish all articles both in Serbian and in English, and was thus an almost perfect resource for testing our tool.

The interest in collections of aligned texts and tools tailored for their search is increasing substantially, primarily due to the growing needs of statistical machine translation. Thus, for example, the OPUS corpus offers freely available parallel corpora in many languages, as well as interfaces for querying the corpus data [Tiedemann, 2009]. Another example of a system that uses parallel corpora for information retrieval is given in [Gravano, 2006].

The HLT group has previously tackled the issue of enhanced cross-lingual search, albeit for individual texts only. Namely, one of the tools developed within the group, dubbed LeXimir, was designed as an integrated environment for various resources, such as morphological e-dictionaries, wordnets, and multilingual proper name databases, which enables, among other things, versatile handling of both monolingual and aligned or comparable texts. LeXimir provides for enhanced querying of aligned texts by using available lexical resources to perform semantic and morphological expansion of queries. The tool was, however, unsuitable for large collections of documents such as multilingual digital libraries of e-journals. Namely, as the search in LeXimir is performed with XPath queries the documents had to be previously loaded in internal computer memory, which is not a feasible request in the case of large collections of texts. In order to realize cross-lingual search for such collections we had to turn to XML databases, which handle storage and indexing of collections of XML files, and enable their search by means of the XQuery language. Thus, the development of a new tool was necessary, one that would extend LeXimir's capabilities to XML databases.

2. The resource

Journal articles from INFOtheca fall in the category of unstructured and semi-structured collections of data. This term generally refers to heterogeneous (different in format, standard and size, etc.) data, of a potentially variable structure, the content of which can grow exponentially. Manipulation of such data, which in general encompass documents, metadata content, user-generated content, RSS feeds, e-mails, geospatial data, etc. asks for an approach that differs from data manipulation in classical relational databases.

Development of XML databases was a natural response to

exponential growth of interchange and storage of data in XML format. A database that manages XML files can be directly integrated into applications (e.g. web services), without the unnecessary slowdown caused by additional layers for interpretation. Native XML databases (abbr. NXD) handle XML documents as their basic logical units, pretty much the same way as relational databases handle rows in tables. Data in NXDs are manipulated only indirectly, by means of database and XML query languages: XQuery, XPath, XML Infoset, DOM and events from SAX 1.0. The NXDs are not expected to substitute existing databases, but rather to offer a tool for development of more robust systems that enable efficient management of XML documents. For the purpose of our research, we have tested two NXD implementations, eXist-db and MarkLogic, and we chose the latter as the more robust and stable platform.

MarkLogic Server 5 (http://www.marklogic.com) is a database for managing, leveraging and delivering unstructured information, developed to meet today's information management requirements. It is designed and optimized to address the common characteristics of unstructured information in terms of handling information that may be human-generated, textual, irregular, hierarchical, de-normalized, time-varying, and/or big. Bibliša, the tool we have built using MarkLogic is aimed for search of document collections consisting of aligned parallel texts converted in TMX (Translation Memory eXchange) format. TMX is an open XML-based standard intended for easier exchange of translation memory data, that is, aligned parallel texts, between tools and translation vendors [TMX, 2005].

TMX uses ISO standards for country and language codes, as well as for date and time. A TMX document consists of a header (metadata describing the aligned texts) and a body, containing a set of translation units (TU) composed of two or more semantically equivalent translation unit variants (TUVs). Each TUV contains the text (one or more sentences or segments) in one of the TMX document languages, where the text in the first TUV is usually in the source language, and the texts in the remaining TUVs are in one or more target languages. Although the order of languages is the same in each TU, there is a TUV attribute xml:lang that denotes the language of the text within the TUV.

The performance of Bibliša was tested on a TMX document collection generated from journal articles published in INFOtheca by another of our tools, dubbed ACIDE [Obradovic et al., 2008]. ACIDE is an integrated development environment for generating aligned parallel texts. It is basically a front-end for two alignment tools developed by LORIA (Laboratoire lorrain de recherche en informatique et ses applications), one for automatic sentence alignment of texts (Xalign, http://led.loria.fr/outils/ALIGN/align.html), and another for alignment visualization and manual correction of alignment errors (Concordancier). ACIDE provides export routines for automatic generation of TMX and HTML versions of parallel texts aligned by the two LORIA tools. The original languages of articles in INFOtheca can be either Serbian or English, thus they switch places as the source and target language. The example in Figure 1 shows a TU from an INFOtheca journal article in English translated into Serbian:

```
<tu>
 <prop type="Domain">Elia et al., 2010, vol. XI:1, ID: 2010.1.2</prop>
 <tuv xml:lang="en" creationid="n7 " creationdate="20110513T151548Z">
  <seg>Abstract: The term "Semantic Web" indicates a set of programs that utilize semantic tags for the information
  retrieval within texts diffused on the Web. </seg>
 </tuv>
 <tuv xml:lang="sr" creationid="n7 " creationdate="20110513T151548Z">
  <seg>Apstrakt: Termin "semantički veb" označava skup programa koji koriste semantičke etikete za pronalaženje
  informacija unutar tekstova rasutih po vebu. </seg>
 </tuv>
</tu>
```

Figure 1. A translation unit (TU) from text in English translated into Serbian

Besides TMX documents generated from articles published bilingually in INFOtheca, our document collection contains metadata in XML format (Figure 2) on all articles and the journal itself, structured as follows:

1. metadata on the journal <JournalCollection> – journal title, ISSN (International Standard Serial Number) and URL of the collection;
2. metadata on the journal issue <Journal> - numeration – number, volume, month and year of publication;
3. metadata on each article <Document>:
   - description metadata – authorship and title proper – author name, affiliation, e-mail, article title and its categorization as well as pagination within the issue, and
   - content metadata – abstract, keywords, and expert classification according to Universal Decimal Classification (UDC).

```
<Document>
 <ID>2010.2.3</ID>
 <UDC>005.923.2:004]:02·026:004.738.52</UDC>
 <Author>
  <ID>1</ID>
  <Name>Paolo·Budroni</Name>
  <Mail>paolo.budroni@univie.ac.at</Mail>
  <Institution·xml:lang="en">Vienna·University,·Library·and·archive·services,·Project·Phaidra</Institution>
  <Institution·xml:lang="sr">Univerzitet·u·Beču,·Biblioteka·i·arhiva,·Projekat·Phaidra</Institution>
 </Author>
 <About·xml:lang="en">
  <Title>Rethinking·How·to·Shape·a·New·Matrix·for·the·Protection·and·Retention·of·Cultural·Heritage</Title>
  <Category>Scientific·paper</Category>
  <URL·xml:lang="en">http://www.zbus.rs/eng/infoteka/dva2010/INFOTHECA_XI_2_December2010_39a-44a.pdf</URL>
 </About>
 <About·xml:lang="sr">
  <Title>Preispitivanje·i·uobličavanje·nove·matrice·zaštite·i·očuvanja·kulturnog·nasleđa</Title>
  <Category>Naučni·rad</Category>
  <URL·xml:lang="sr">http://www.zbus.rs/cir/infoteka/dva2010/INFOTHECA_XI_2_December2010_37-43.pdf</URL>
 </About>
```

Figure 2. An example of metadata

All metadata, except language independent data, such as the numeration metadata (<ID>, <Number>, <Volume>, <Month>, <Year>), the <UDC> and <Mail>, are entered in both languages (Serbian and English), using the attribute xml:lang to denote the language of the content (see Figure 2). These metadata can be used for refinement of "full text" document search, as well as for additional forms of searching and browsing, as illustrated in Section 5.

The current collection consists of 44 TMX documents representing aligned parallel versions of INFOtheca articles published in the period 2007-2011. Both Serbian and English version of journal articles were manually preprocessed and then automatically aligned, and the alignments manually corrected using ACIDE. Table 1 shows the total, minimum, maximum and average length of articles (in words and sentences), given separately for Serbian and English.

Sr	**Tokens**	**Types**	**Words**	**Word types**	**Sent ences**
total	200,694	68,501	159,031	65,488	7,986
min	746	369	609	342	27
max	12,214	3,260	8,440	3,105	414
avg	4,561	1,557	3,614	1,488	182
En	**Tokens**	**Types**	**Words**	**Word types**	**Sent ences**
total	220,120	50,747	178,269	47,719	7,986
min	838	336	718	310	27
max	13,919	2,300	10,005	2,183	414
avg	5,003	1,153	4,052	1,085	182

Table 1: Statistics for Serbian and English documents

3. Software implementation

The Bibliša tool operates within a complex system composed of several modules as depicted in Figure 3. Targeted at textual resources in the form of collections of TMX documents and the corresponding metadata, the system has at its disposal several other lexical resources, such as morphological e-dictionaries. Together with the system of rules for compound inflection, finite automata and transducers, these dictionaries represent the basis for morphological expansion of queries. As for semantic and bilingual expansion, the system relies on wordnets (Serbian and English at present) and a bilingual English/Serbian dictionary of Library and Information Science technology (see Section 4).

Bibliša's user formulates the initial query in the form of one or more simple or multiword keywords, which is then forwarded for further morphological and semantic expansion, depending on the user's preferences. Query expansion is basically handled by a web service (wsQueryExpand.asmx) and a web application (VebRanka). VebRanka, a workstation for query expansion, was developed as a web extension of our multipurpose tool LeXimir [Stankovic et al., 2011], with the aim of managing the complex task of query expansion on the web. The web service, which receives the query from the web application, invokes the function library (LeXimirCore), a component of LeXimir. The library uses lexical resources and Unitex (http://igm.univ-mlv.fr/~unitex) routines to perform an expansion of the query according to user specifications, and returns the result to the web service, which in its turn returns it to the web application that invoked him. When the expanded query reaches Bibliša, the tool forwards it to the document collection in the form of a XQuery. In the excerpt from XQuery code shown in the following example the variable $x creates a collection of "tu" XML nodes (elements), which can subsequently be queried, yielding the results obtained by $query tagged with :

```
for $x in cts:search(fn:doc()//tu,$query)
return
cts:highlight($x, $query, <em>{$cts:text}</em>)
```

The results, a set of aligned concordances, are formatted and presented to the user. The concordances are preceded by information identifying the document they originate from, and a link to summary metadata for this document in both languages .

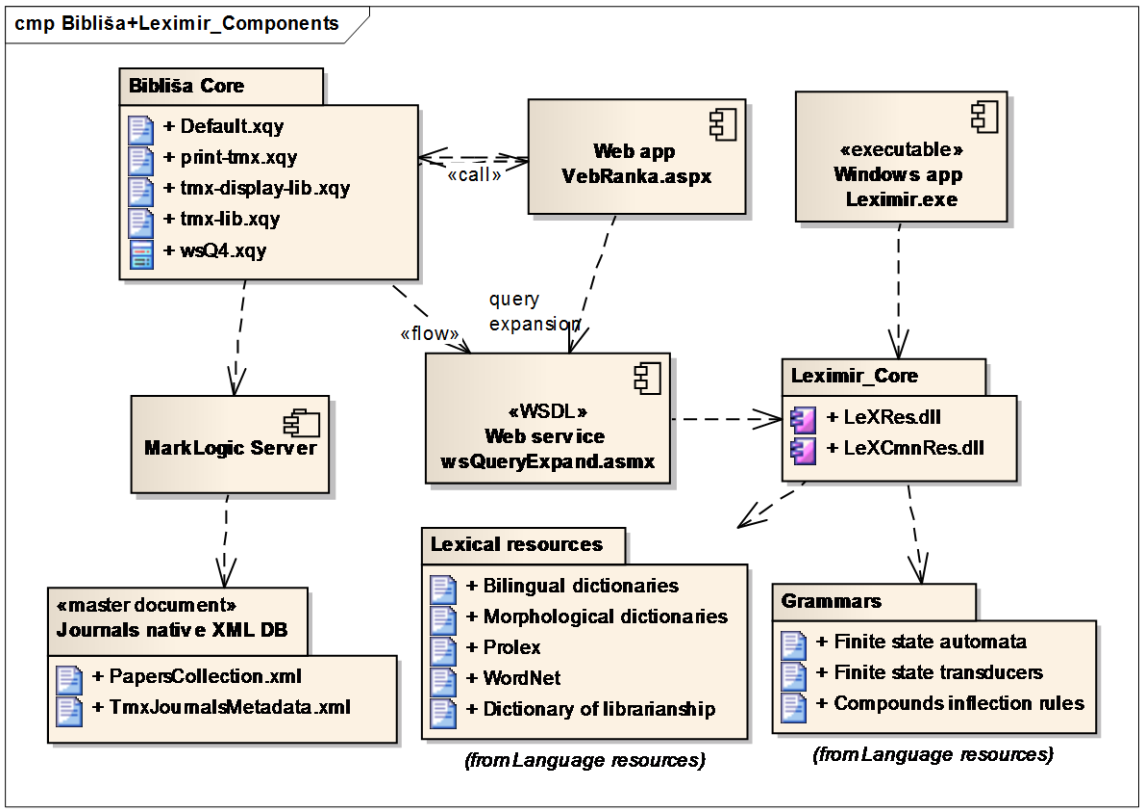


Figure 3. Bibliša+LeXimir UML component model

4. Supporting resources

Three types of lexical resources are used for the expansion of queries submitted to our collection of documents. The most important resources are Serbian morphological dictionaries of simple words and multi-word units [Krstev, 2008]. These comprehensive resources were developed and are being mainly used within two corpus processing systems: Unitex and Nooj. However, Unitex standalone routines enable the usage of morphological dictionaries developed under Unitex in other environments. The dictionaries are used to generate all inflective forms of query keywords, thus improving the system recall without negative effects on precision. Moreover, for multi-word units not found in dictionaries, there exists a rule-based strategy, which attempts to recognize syntactic structure of the multi-word and how it should be inflected [Stankovic et al., 2011].

Another type of resources used by Bibliša for bilingual and semantic expansion of queries are wordnets. At present we use the Princeton English Wordnet (PWN), version 2.0, as well as the Serbian Wordnet (SrpWN — http://korpus.matf.bg.ac.rs/SrpWN), initially developed in the scope of the BalkaNet project [Tufiș, 2004] and subsequently enhanced and upgraded. Most of the synstes in Serbian Wordnet are aligned with PWN synsets via the Interlingual Index, with the exception of Serbian specific synsets that are not available in PWN. The Interlingual Index enables bilingual query expansion, while synonymous relations expressed by synsets themselves enable semantic expansion. At present, we do not use other semantic relations available in both Wordnets.

Finally, Bibliša also uses the "Dictionary of librarianship: English-Serbian and Serbian-English" that is available online at http://btr.nb.rs. The production of this dictionary started in 2001 at the National Library of Serbia's, with the aim of producing a domain specific dictionary in various formats and on different media. This endeavour was the continuation and upgrade of the international project "Multilingual dictionary of library terminology" that started in the 1990s, with the aim of collecting and presenting the librarianship terminology for 16 European languages [Kovačević et al., 2004].

The online version of the Dictionary currently contains 23 400 terms – 11 300 in English and 12 100 in Serbian, 910 definition or annotation terms which belong to library standards, and 2 200 acronyms of international and national entities. It covers both the ekavian and ijekavian Serbian dialects. For our purposes we used the Excel format of the dictionary, which we transformed into a MS SQL Server 2008 relational database. Due to the complexity of the Excel format the transformation turned out to be a rather complicated procedure. Nevertheless, we managed to preserve all important features of the original dictionary: the relations between translational equivalents, the synonymous relations within each specific language, and for Serbian, both the ekavian and ijekavian. Finally, this newly obtained resource was normalized in order to make it more suitable for the search of collections of aligned texts.

5. User interface

The user can search the INFOtheca collection in two different ways. A typical search, yielding a set of aligned concordances is a full text search based on a query, and we will discuss this type of search shortly. However, Bibliša offers also the possibility of a monolingual search of metadata, associated with every article in the collection and described in more detail in Section 2. When performing such a search the user can make use of a form with predefined fields for the most commonly used data: author's name, words from an article title, year of publication, and article keywords. The search can be performed either in the English or the Serbian part of the collection (Figure 4).

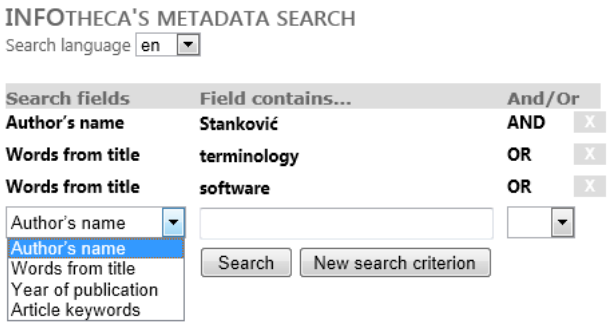


Figure 4. Interface for metadata search

As a result, a list of all articles in the language of the query that fulfill the required conditions is produced with all available metadata. In addition to that, for each article, links are offered to the full text of the article in .pdf format (residing on the official site of the INFOtheca journal) as well as the entire aligned parallel text of the article in .html format.

More powerful is the full-text search (Figure 5). The user initiates this search by submitting a keyword in English or in Serbian and choosing the resources to be used for its semantic expansion by checking appropriate boxes (WordNet and/or Dictionary of librarianship). The system responds with several editable lists of keywords depending on what is found in resources chosen for expansion. The user can edit these lists by deleting some keywords or adding new ones. For example, if the user submitted *biblioteka*, as the query keyword in Serbian, and asked for semantic expansion using wordnets, Bibliša would use the Serbian wordnet to connect to the English wordnet. The English wordnet would then expand the query by the following keywords: *library*, *program library*, *subroutine library*, *bookcase* and *bibliotheca*. It is quite probable that the user might want to remove some of these keywords, e.g. *bookcase*.

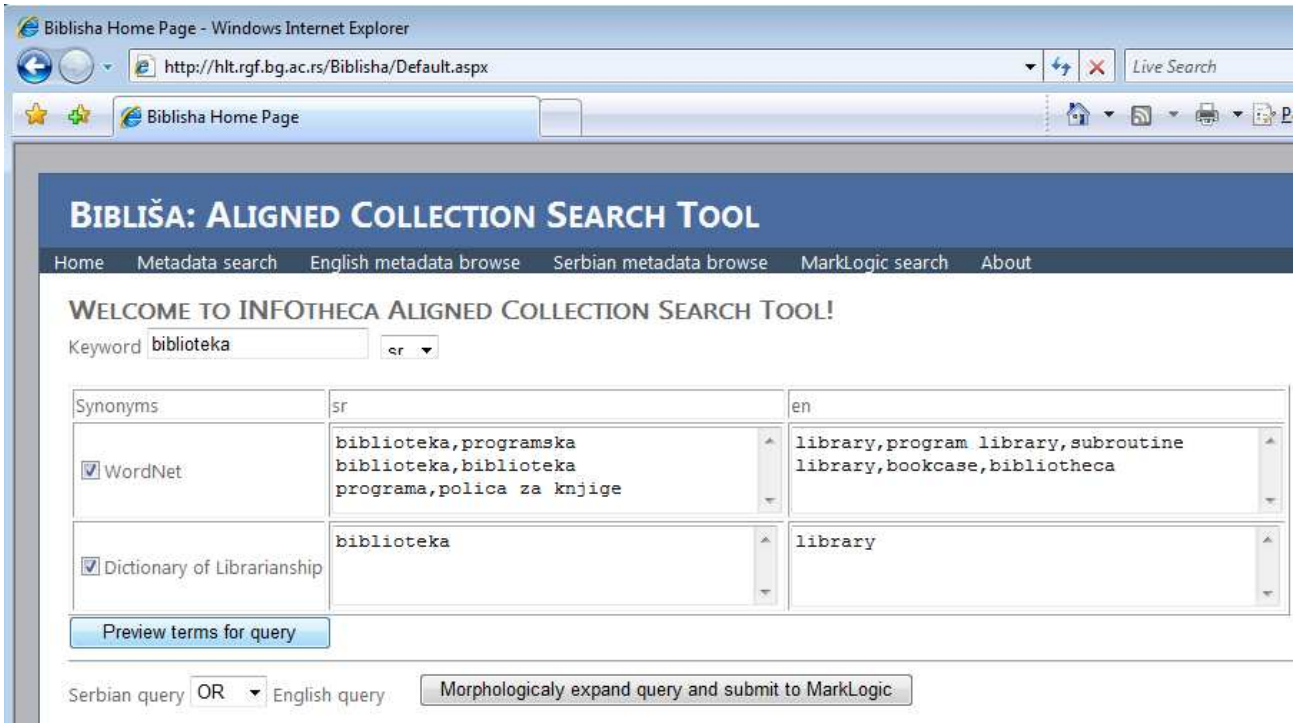


Figure 5. Interface for query expansion and submit

After semantic expansion, the set of Serbian keywords chosen and verified by the user is expanded morphologically. Namely, MarkLogic supports stemming in several languages including English but not Serbian. Thus, the morphological forms of English keywords are taken care of by MarkLogic's stemming capability. As for Serbian, Bibliša must expand the initial query with all morphological forms of the keywords using the available Serbian resources (SrpRec) (see Section 4).

The fully, semantically and morphologically expanded query is then used to search through TUVs taking into account the value of the xml:lang attribute. Namely,

English keywords will search only through TUVs with xml:lang attribute set to *en*, while Serbian keywords will search through TUVs with xml:lang attribute set to *sr*. For instance, if the search is initiated with the English keyword *online*, and the Dictionary of librarianship is selected as the resource for semantic and bilingual expansion, *on-line* will be added as another English keyword, and *onlajn* and *u mreži* as Serbian keywords. A concurrent search with these two lists will reveal that both English terms: *online* and *on-line* are used in Serbian texts as well. However, they will not be looked for and hence not recognized and highlighted in the results.

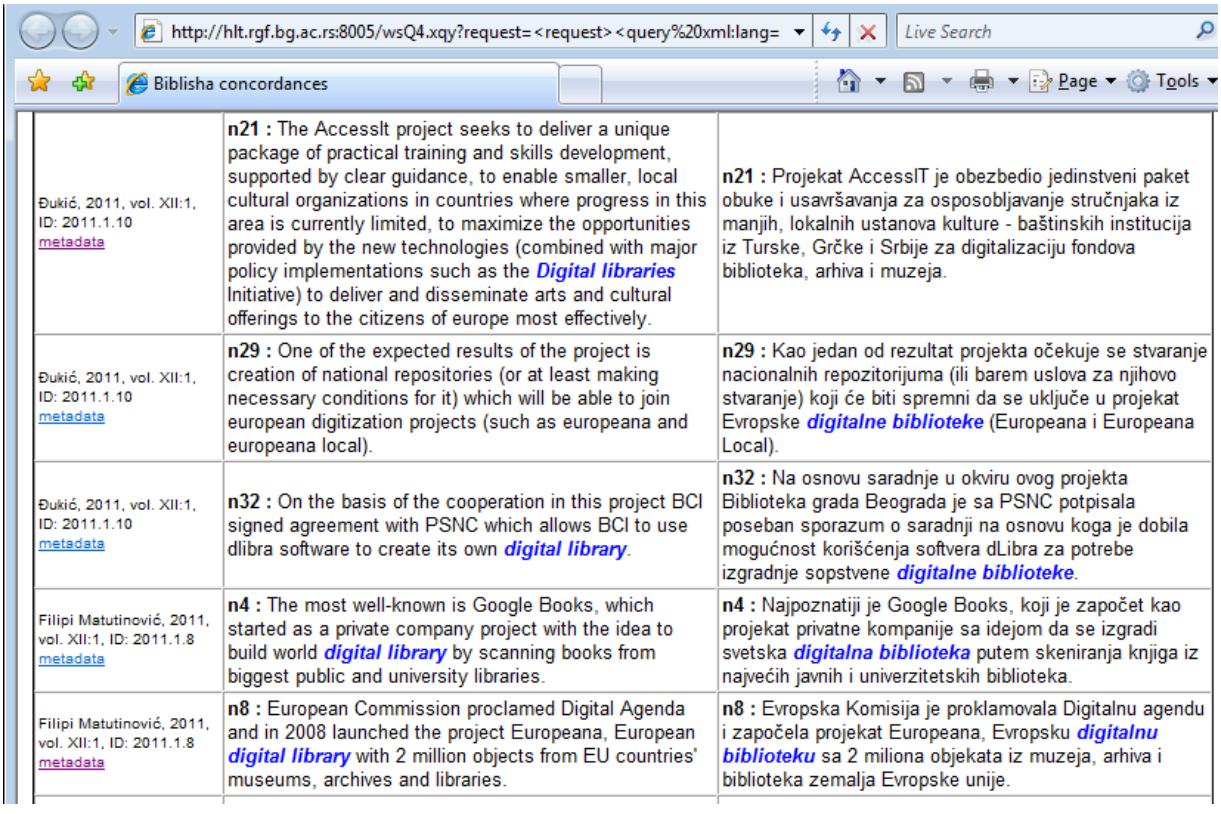
To illustrate morphological expansion of a query in Bibliša we will use the keyword *digitalna biblioteka* (*digital library*). As we have already mentioned, MarkLogic stemming capability will take care of the morphological forms of the compound in English, while Bibliša needs to expand the query with all morphological forms of the compound in Serbian. The expanded query generated by Bibliša thus consists of several Serbian entries and one English entry:

```
<request>
   <query xml:lang='sr'>digitalna biblioteka</query>
   <query xml:lang='sr'>digitalnih biblioteka</query>
   <query xml:lang='sr'>digitalne biblioteke</query>
   <query xml:lang='sr'>digitalnim bibliotekama</query>
   ....
   <query xml:lang='en'>digital library</query>
</request>
```

Morphological expansion is performed automatically and is not transparent to the end user. Part of the XQuery code that follows is used for transforming the expanded query to the proper XQuery form:

```
cts:or-query((cts:word-query("digitalna biblioteka",
("stemmed", "lang= sr") ).
cts:or-query((cts:word-query("digitalnih biblioteka",
("stemmed", "lang= sr") ).
cts:or-query( (cts:word-query("digitalne biblioteke",
("stemmed", "lang= sr") ),
cts:or-query( (cts:word-query("digitalnim bibliotekama",
("stemmed", "lang= sr") ),
.....
                    cts:word-query(digital library,
("stemmed", "lang= en") ) ))  )))
```

Starting from the initial query, processed as outlined, the system finally generates aligned concordances in which all retrieved keywords are highlighted in both languages. Each concordance line is preceded by an identification of the document it originates from. Within this identification is also a link to full metadata of the document. A part of concordances for the initial query *digitalna biblioteka* is presented in Figure 6.



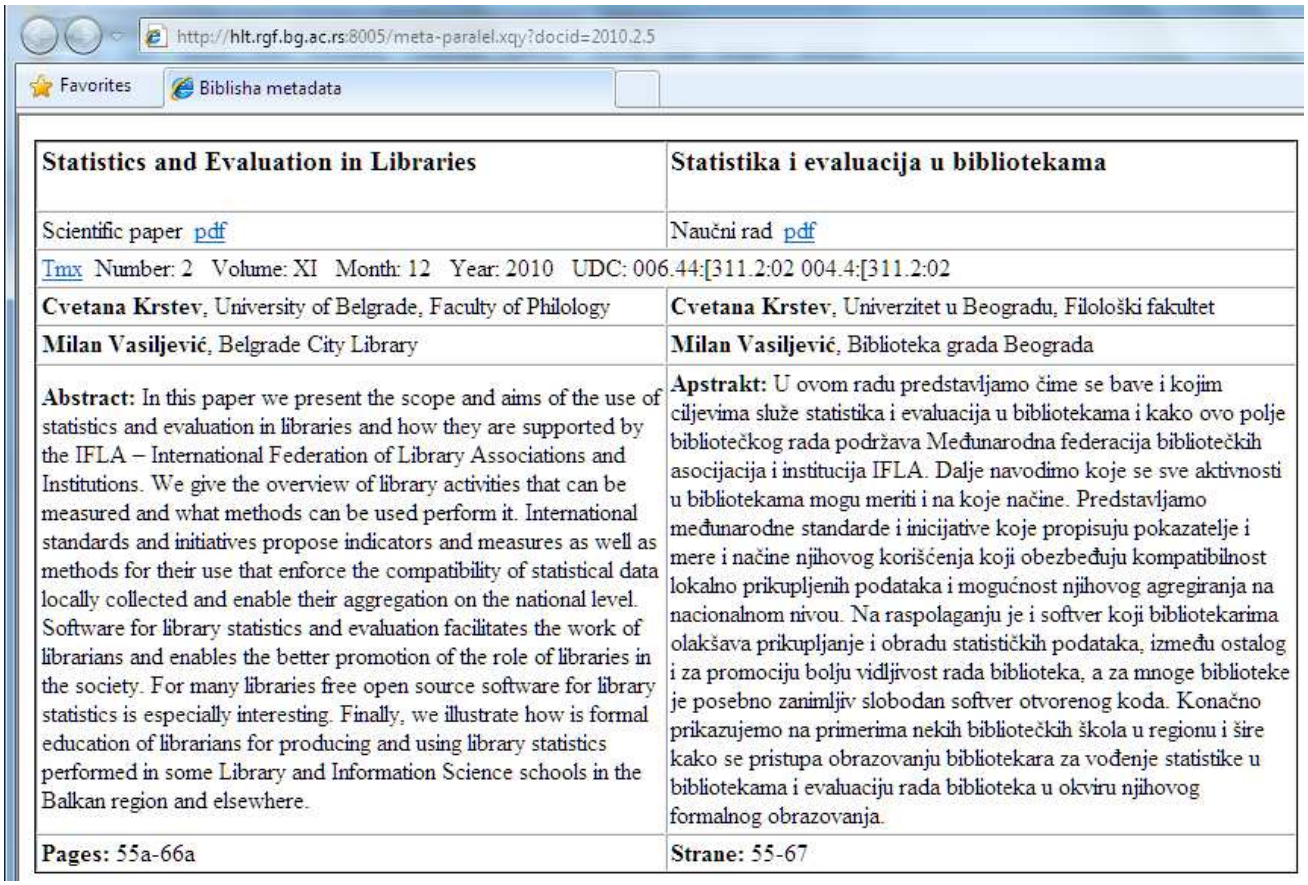
Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n21 :** The AccessIt project seeks to deliver a unique package of practical training and skills development, supported by clear guidance, to enable smaller, local cultural organizations in countries where progress in this area is currently limited, to maximize the opportunities provided by the new technologies (combined with major policy implementations such as the ***Digital libraries*** Initiative) to deliver and disseminate arts and cultural offerings to the citizens of europe most effectively.	**n21 :** Projekat AccessIT je obezbedio jedinstveni paket obuke i usavršavanja za osposobljavanje stručnjaka iz manjih, lokalnih ustanova kulture - baštinskih institucija iz Turske, Grčke i Srbije za digitalizaciju fondova biblioteka, arhiva i muzeja.
Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n29 :** One of the expected results of the project is creation of national repositories (or at least making necessary conditions for it) which will be able to join european digitization projects (such as europeana and europeana local).	**n29 :** Kao jedan od rezultat projekta očekuje se stvaranje nacionalnih repozitorijuma (ili barem uslova za njihovo stvaranje) koji će biti spremni da se uključe u projekat Evropske ***digitalne biblioteke*** (Europeana i Europeana Local).
Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n32 :** On the basis of the cooperation in this project BCI signed agreement with PSNC which allows BCI to use dlibra software to create its own ***digital library***.	**n32 :** Na osnovu saradnje u okviru ovog projekta Biblioteka grada Beograda je sa PSNC potpisala poseban sporazum o saradnji na osnovu koga je dobila mogućnost korišćenja softvera dLibra za potrebe izgradnje sopstvene ***digitalne biblioteke***.
Filipi Matutinović, 2011, vol. XII:1, ID: 2011.1.8 metadata	**n4 :** The most well-known is Google Books, which started as a private company project with the idea to build world ***digital library*** by scanning books from biggest public and university libraries.	**n4 :** Najpoznatiji je Google Books, koji je započet kao projekat privatne kompanije sa idejom da se izgradi svetska ***digitalna biblioteka*** putem skeniranja knjiga iz najvećih javnih i univerzitetskih biblioteka.
Filipi Matutinović, 2011, vol. XII:1, ID: 2011.1.8 metadata	**n8 :** European Commission proclamed Digital Agenda and in 2008 launched the project Europeana, European ***digital library*** with 2 million objects from EU countries' museums, archives and libraries.	**n8 :** Evropska Komisija je proklamovala Digitalnu agendu i započela projekat Europeana, Evropsku ***digitalnu biblioteku*** sa 2 miliona objekata iz muzeja, arhiva i biblioteka zemalja Evropske unije.

Figure 6. Concordances for the initial query *digitalna biblioteka*

If the user is interested in metadata, or the full text of a specific document, he/she can obtain it by clicking on the "metadata" link shown in the leftmost column in Figure 6. This opens the metadata window (Figure 7) showing the document metadata in both languages and offering further links to full texts of the document in both languages (pdf), as well as to the full parallel aligned text in TMX format (tmx).

The user can choose between two options for retrieving concordances: either only concordances where one or more keywords were found in both languages (AND), or concordances in which one or more keywords were found in at least one of them (OR). By default, English and Serbian keyword lists are used as if connected by the "OR" operator. This way, the user can obtain text segments in which either the Serbian or the English term was translated in an unexpected way. For instance, for the English term *browser* the Dictionary of library and information science terminology offers the Serbian translations *pretraživač*. The system retrieves 14 segments none of which contain both the English and the Serbian highlighted term. Namely, English *browser*, referring to web browser program is never translated with either *pretraživač*, while retrieved occurrences of *pretraživač* never refer to a browser.

However, in some specific cases it can be useful to search with English and Serbian keyword lists connected by the "AND" operator. This type of search retrieves only segments that contain both the English and the Serbian term, thus excluding unwanted hits resulting from the ambiguity of terms. For instance, when initiating a search with the Serbian term *novine*, and expanding it with both the Wordnets and the Dictionary of librarianship the following keyword lists are obtained: *newspaper*, *paper* and *press* for English and *dnevne novine*, *dnevnik*, *novine*, and *štampa* for Serbian. If the search is performed with the "OR" option, 238 resulting aligned segments are obtained. However, the English term *paper* is highly ambiguous and as a consequence, numerous retrieved results containing it are false. By performing the search with the "AND" option the number of concordances is reduced to 29, and all false results disappear.



Statistics and Evaluation in Libraries	**Statistika i evaluacija u bibliotekama**
Scientific paper pdf	Naučni rad pdf
Tmx Number: 2 Volume: XI Month: 12 Year: 2010 UDC: 006.44:[311.2:02 004.4:[311.2:02	
Cvetana Krstev, University of Belgrade, Faculty of Philology	**Cvetana Krstev**, Univerzitet u Beogradu, Filološki fakultet
Milan Vasiljević, Belgrade City Library	**Milan Vasiljević**, Biblioteka grada Beograda
Abstract: In this paper we present the scope and aims of the use of statistics and evaluation in libraries and how they are supported by the IFLA – International Federation of Library Associations and Institutions. We give the overview of library activities that can be measured and what methods can be used perform it. International standards and initiatives propose indicators and measures as well as methods for their use that enforce the compatibility of statistical data locally collected and enable their aggregation on the national level. Software for library statistics and evaluation facilitates the work of librarians and enables the better promotion of the role of libraries in the society. For many libraries free open source software for library statistics is especially interesting. Finally, we illustrate how is formal education of librarians for producing and using library statistics performed in some Library and Information Science schools in the Balkan region and elsewhere.	**Apstrakt:** U ovom radu predstavljamo čime se bave i kojim ciljevima služe statistika i evaluacija u bibliotekama i kako ovo polje bibliotečkog rada podržava Međunarodna federacija bibliotečkih asocijacija i institucija IFLA. Dalje navodimo koje se sve aktivnosti u bibliotekama mogu meriti i na koje načine. Predstavljamo međunarodne standarde i inicijative koje propisuju pokazatelje i mere i načine njihovog korišćenja koji obezbeđuju kompatibilnost lokalno prikupljenih podataka i mogućnost njihovog agregiranja na nacionalnom nivou. Na raspolaganju je i softver koji bibliotekarima olakšava prikupljanje i obradu statističkih podataka, između ostalog i za promociju bolju vidljivost rada biblioteka, a za mnoge biblioteke je posebno zanimljiv slobodan softver otvorenog koda. Konačno prikazujemo na primerima nekih bibliotečkih škola u regionu i šire kako se pristupa obrazovanju bibliotekara za vođenje statistike u bibliotekama i evaluaciju rada biblioteka u okviru njihovog formalnog obrazovanja.
Pages: 55a-66a	**Strane:** 55-67

Figure 7. Metadata window with links to full texts (pdf) and to full parallel aligned text (tmx)

6. Future work

The performance of Bibliša, our tool enabling enhanced search of multilingual digital libraries of e-journals, has been tested on a TMX document collection generated from 44 articles and their translations published in the journal INFOtecha. Although the tool is still under development, and hence not fully operational, the results of the initial tests were very encouraging.

In the nearest future the interface of the tool will be further enhanced, and its functionalities improved, taking especially into account the comments and suggestions of future users. Our aim is to enable Bibliša to become more than a tool for powerful full-text and metadata search. We believe that with some additional modules we can make it a useful translator's aid in reviewing terminology used in context. Furthermore, it could be used to improve existing terminology in the domain of Library and Information Science. For example, the analysis of concordances obtained for *browser/pretraživač* (Section 5) revealed that the correct translation for *browser* in Serbian should be *prelistač*, whereas the correct translation for *pretraživač* in English should be *search engine*. Hence, the Dictionary of librarianship should be amended accordingly. In this context, an additional module is planned, which would enable privileged users to enhance and correct the dictionaries used in query expansion on the basis of search results.

So far, we have used wordnets and the Dictionary of librarianship as supporting multilingual resources. In the future we plan to expand the set of these resources with Prolex-a, a multilingual ontology of proper names (http://www.cnrtl.fr/lexiques/prolex) and possibly other multilingual resources as well. On the other hand, we also plan to exploit more the possibilities for semantic expansion offered by existing resources, especially the abundance of semantic relations that exist in wordnets besides synonymous. For example, the use of derivational relation would enable us to expand *novine* (*newspaper* – noun) with *novinski* (*newspaper* – adjective) thus improving recall in the case when the user opts for the "AND" search (Section 5).

7. Acknowledgements

Preprocessing of texts and correction of alignment were done by Jelena Andonovski, PhD student at the Faculty of Philology. This research was supported by the CESAR (Central and South-East European Resources) project (ICT PSP programme, no. 271022) and by Serbian Ministry of Education and Science under the grant #III 47003.

References

Gravano, L. Nezinger, M.H. (2006). Systems and Methods for Using Anchor Text as Parallel Corpora for Cross-Language Information Retrieval - US Patent 7,146,358 B1 - Google Patents.

Kovačević, Lj., Injac, V., Begenišić, D. (2004). Bibliotekarski terminološki rečnik - englesko-srpski, srpsko-engleski, Beograd: Narodna biblioteka Srbije.

Krstev, C. (2008). Processing of Serbian – Automata, Texts and Electronic dictionaries. Belgrade: Faculty of Philology, University of Belgrade.

Obradović, I., Stanković, R., Utvić, M. (2008). An Integrated Environment for Development of Parallel Corpora (in Serbian). In: Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen (pp. 563-578), B. Tošović (Ed.). Berlin: LitVerlag.

Stanković, R., Obradović, I., Krstev, C., Vitas, D. (2011), "Production of morphological dictionaries of multi-word units using a multipurpose tool", In: Proceedings of the Computational *Linguistics-Applications Conference*, October 17–19, 2011. Jachranka, Poland (pp. 77-84), K. Jassem, P. W. Fuglewicz, M. Piasecki and A. Przepiórkowski (eds.), Polish Information Processing Society, ISBN 978-83-60810-47-7

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Recent Advances in Natural Language Processing (vol. 5) (pp 237-248), N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.). John Benjamins, Amsterdam/Philadelphia.

TMX 1.4b Specification. (2005). http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

Tufiş, D. (ed.) (2004). Romanian Journal on Information Science and Technology. Special Issue on BalkaNet, vol. 7. Romanian Academy, 248 p. ISSN 1453-8245.