A.A. Baptista et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-065-9-62

Ontology-Based Search and Document Retrieval in a Digital Library with Folk Songs

Maria NISHEVA-PAVLOVA[1] and Pavel PAVLOV
Faculty of Mathematics and Informatics, Sofia University

Abstract. The paper discusses some aspects of a work in progress aimed at the development of technologies for digitization of Bulgarian folk music and building a digital library with Bulgarian folk songs presented with their music, texts and notes. This library provides digital preservation of the sound recordings, lyrics and notations of more than 1000 Bulgarian folk songs and tools for various types of search and analysis of the available resources. The presentation is focused on the subject ontology especially developed for the occasion and its application in the implementation of a tool for semantics oriented search in the lyrics of songs.

Keywords. digital library, metadata, ontology, search engine, document retrieval

Introduction

Traditional Information Retrieval technology is almost entirely based on the occurrence of words in documents and therefore it may be characterized as keywords-based. Search engines augment this in the context of the Web with information about the hyperlink structure of the Web. In the typical cases, the user provides the search engine with a word or phrase about which he/she is trying to gather or research information. The kinds of queries a keyword-based system can accept are quite limited and semantically poor. Moreover, simple keyword-based search usually returns too many results which have to be additionally filtered somehow.

Semantic Search [1] attempts to augment and improve traditional search results. Ontologies play a key role in this kind of search. An ontology, by definition, represents a formal model of the common interpretation of the entities and relationships in a domain of interest. Therefore, ontologies should be widely used in digital library systems. In particular, at least three types of ontologies have been identified as applicable in a specific type of digital libraries – the so-called *Semantic Digital Libraries* [2,3,4]: bibliographic ontologies, ontologies for content structures (or subject ontologies, according to the term used in this paper), community-aware ontologies. Subject ontologies are useful for supporting the semantic annotation of all types of library resources. They also play the role of knowledge sources which define the meaning of most domain concepts, their hierarchy, properties and relationships.

[1] Corresponding Author: Maria Nisheva-Pavlova, Faculty of Mathematics and Informatics – Sofia University, 5 James Bourchier blvd., 1164 Sofia, Bulgaria; E-mail: marian@fmi.uni-sofia.bg.

The paper discusses some results of the activities within an ongoing project aimed at the development of technologies for digitization of Bulgarian folk music and building a semantic digital library (named DjDL) with Bulgarian folk songs presented with their notes, text and music. DjDL is intended to serve as a platform for digital preservation of the sound recordings, lyrics and notations of a significant number of Bulgarian folk songs and to provide adequate access to them. The emphasis of the presentation falls on the provided tool for semantic (ontology-based) search in the lyrics of songs.

1. An Overview of DjDL

Currently DjDL contains a collection of over 1000 digital objects which represent a part of the unpublished archive manuscripts of Prof. Todor Dzhidzhev including folk songs from the Thracia region of Bulgaria.

DjDL has the typical architecture of an academic digital library with heterogeneous resources. Its functional structure includes six main components:

- a metadata catalogue;
- a repository;
- a subject ontology;
- a search engine;
- a module implementing the library functionality;
- an interface module.

The library catalogue consists of short descriptions (in XML format) of the particular folk songs included in the repository. These descriptions contain various types of metadata, for example: the title of the song, the song genre in accordance with different classification schemes (e.g. according to the typical time and space of performance, the thematic focus(es), the context of performance, etc.), the region of folk dialect, the informant (the person who conveyed the song to folklorists), the folklorist who gathered the song, the singer(s), the date and place of record, etc. More accurately, each catalogue entry contains the text (i.e., the lyrics) of a particular song accompanied with the corresponding metadata.

The repository of DjDL contains heterogeneous resources of four types [5,6]:

- lyrics of songs (in PDF format);
- notations of songs (in LilyPond format [7] and properly visualized in PDF format);
- musical (MP3) files with the authentic performances of the songs – as far as such exist in the archives;
- musical (MIDI) files generated with the use of LilyPond from the notations of the songs.

The subject ontology includes concepts of different areas related to the contents of the lyrics of songs, with description of their properties and different kinds of relationships among them. It plays a significant role in the implementation of the full functionality of the search engine.

The purpose of the search engine is to provide adequate access to the complete palette of resources stored in DjDL.

The library functionality and the user interface of DjDL are designed in accordance with the expected requirements of the typical users of the library. The

interface module provides adequate online access to the library resources and supporting software tools.

2. Subject Ontology

The subject ontology describes a proper amount of domain knowledge (with definitions of the main concepts, their properties/relationships and representative instances) that has been used to build the catalogue descriptions and to process the user queries. It consists of several interrelated subontologies needed by the search engine of DjDL and developed especially for the occasion:

- ontology of folk songs – includes various genre classifications of folk songs (by their thematic focus – historical, mythical, etc.; by the context of performance – Christmas folk songs, harvest songs, etc.; by their cultural functions – blessing, oath, wooing, etc.);
- ontology of family and manner of life;
- ontology of impressive events and natural phenomena;
- ontology of social phenomena and relationships;
- ontology of historic events;
- ontology of disasters;
- ontology of feasts;
- ontology of traditions and rites;
- ontology of blessings and curses;
- ontology of mythical creatures and demons;
- ontology of administrative division – combines the current administrative division of Bulgaria with the one from the beginning of XX century.

Most classes of the subject ontology are constructed as defined OWL classes, by means of necessary and sufficient conditions defined in terms of proper restrictions on certain properties (see e.g. Figure 1).

The properties “form” and “synonym” provide the search engine with suitable grammatical forms and synonyms of the terms used as names of ontology classes.

3. Patterns and Rules

The folklore lyrics uses lots of similes, metaphors, idioms and other sophisticated or language-dependent stylistic devices. For that reason, it is expedient to accompany the use of proper ontologies with other Artificial Intelligence tools to provide more adequate support for the semantic search.

In this sense we direct our attention to the design and use of proper patterns of typical stylistic or thematic constructs which could be matched with relatively large parts of the texts of folklore songs. We call them *concept search patterns*. For instance, we have already defined a number of search patterns of constructs standing for “unfaithfulness”, “jealousy”, “discontent”, “sedition” etc. as well as the corresponding pattern matching rules and recently make various experiments with them.

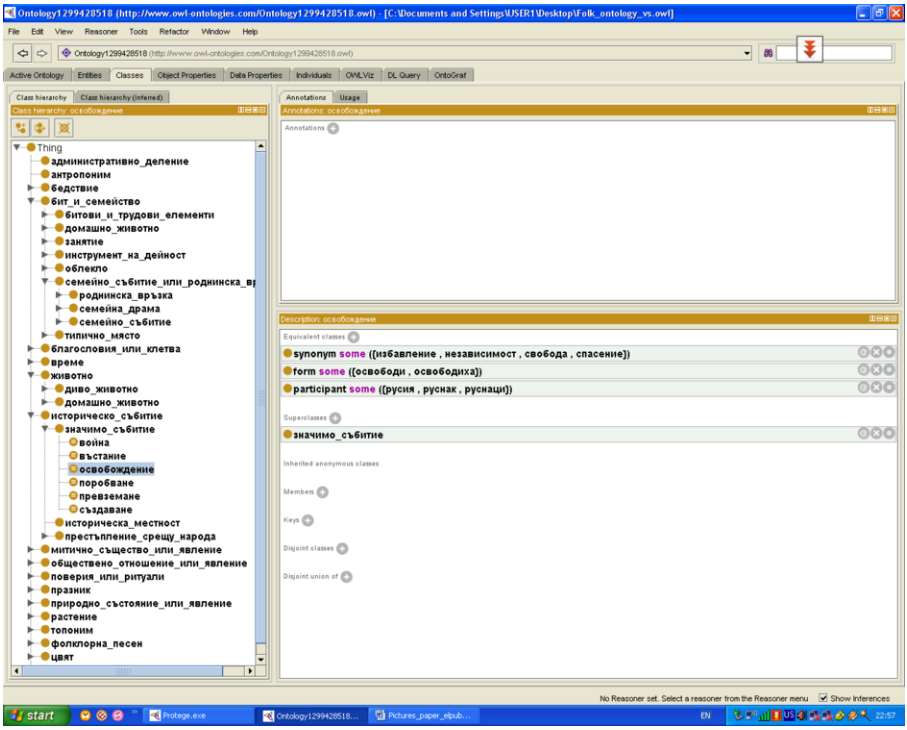

Figure 1. Part of the subject ontology.

Let us consider for example the definition of a pattern for the concept “love infidelity” (“любовна изневяра” in Bulgarian) which is oftentimes used in Bulgarian folk songs. Multiform phrases have been used in the lyrics of songs to express (real or possible or future) love infidelity, e.g. “Та друго либе залиби”, “Друго любе жъ залюбя”, etc.

An essential part of them match the pattern

<< друг$? <любим_а> $? зал?б? >>

Here “?” and “$?” are used as wildcard symbols (the question mark “?” matches any letter at the corresponding position and the symbol “$?” matches any corresponding sequence of zero or more letters) and the angle brackets “< >” enclose the name of an ontology concept (the concept “любим_а“ in the subject ontology means “beloved” in English).

4. Functionalities of the Search Engine

The search engine of DjDL supports two main types of search: keywords-based and semantic (ontology-based) search. The design of this search engine is based on some former results of the authors [8,9] and some ideas from [1] and [10]. Its current version realizes some facilities for search in the catalogue metadata and the lyrics of songs only. The functionalities supported at present were specified after a careful study of the

requirements of the typical user groups (specialists and researchers in ethnomusicology and verbal folklore, philologists, etc.).

The user queries define restrictions on the values of certain metadata attributes and/or the texts of the required folk songs. The search procedure consists of some pattern matching activities in which the catalogue descriptions containing metadata and lyrics of songs are examined one by one and those of them having a specific set of element values that match the corresponding components of the user query, are marked in order to form the search result.

The matching process within the keywords-based search consists in testing the appropriate sources for equality.

During the construction of a query for keywords-based search, the user is asked to indicate the search source(s) – search in the lyrics of songs, search in the metadata or combined search in the lyrics of songs and catalogue metadata. One can define a search query consisting of an arbitrary number of words or phrases as well as specify proper logical connectives between them: conjunction (and) or disjunction (or). Negation (not) is also allowed as a unary operator indicating that the negated word or phrase should not appear in the searched text. As a result of the user query processing, a list of links to the discovered catalogue files with metadata and lyrics of songs has been properly displayed. This list may be ordered according to various criteria.

Here are some typical examples of queries for keywords-based search:

- search and retrieval of songs whose lyrics contain specific words or phrases;
- search (and retrieval) of songs with distinct thematic focus or context of performance;
- search of songs performed by a given singer;
- search of songs performed in a particular settlement;
- search of songs performed by singers from a given place.

The semantic (ontology-based) search tool of DjDL is aimed at the provision of a set of additional facilities for augmentation and refinement (automatic reformulation according to the available explicit domain knowledge) of the queries for keywords-based search.

The augmentation of the user query is based on proper utilization of the subject ontology. First of all, an exhaustive breadth-first search in the graph representing the "is-a" concept hierarchy described by the subject ontology is performed, starting from the node which corresponds to the original user query. The names of the visited nodes that are in fact the respective more specific concepts included in the ontology, are added to the one given by the user. The resulting list of concepts if properly visualized and placed at user's disposal for further refinement (see Figure 2).

Within the next step of query expansion, the search engine adds to the newly constructed set of queries some synonyms and derivatives of the main terms found as their relevant properties in the subject ontology. The corresponding property values from the definitions of all concepts included by that time in the expanded user query and the existing instances of these concepts are added to the query as well.

Thus the user query is augmented as far as possible in terms of the subject ontology and in fact it has the shape of a disjunction of all included forms of concepts and instance names. In this form the resulting query is ready for further refinement (see e.g. Figure 3) and processing.

As example queries for ontology-based search, being of interest for folklorists (according to [11]), that can be executed by the search engine of DjDL, we could indicate the queries for search and retrieval of:

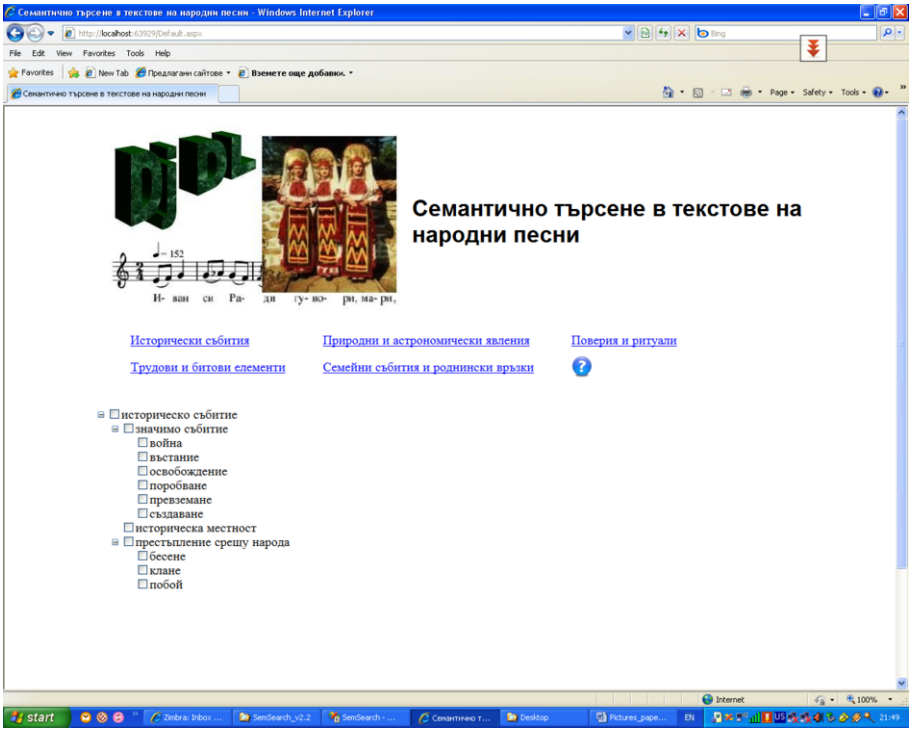

Figure 2. Construction of a user query for ontology-based search (step 1).

- songs devoted to (or mentioning) historic events or important social phenomena;
- songs in which exciting natural or astronomical phenomena are described or mentioned;
- songs in which typical (or typical for a certain region) folk beliefs are described;
- songs in which elements of country work and life are described or mentioned;
- songs in which significant family events or human relations are mentioned.

The search engine provides also some facilities for processing of user queries presuming examination of equality or inequality. For example, it is possible to formulate and execute queries for search of:

- songs performed alone/in a group;
- songs performed by men/women only;
- songs performed by one and the same singer;
- songs performed by singers, born in one and the same settlement or region;
- songs performed in a specific region (grouped by regions of performance);

- songs performed in settlements to the west/east/north/south of a specific settlement/region.

The current version of the search engine of DjDL is provided by a very small number of concept search patterns, therefore their application is still quite limited. When appropriate, the pattern matcher performs the last step of the execution of queries for ontology-based search.

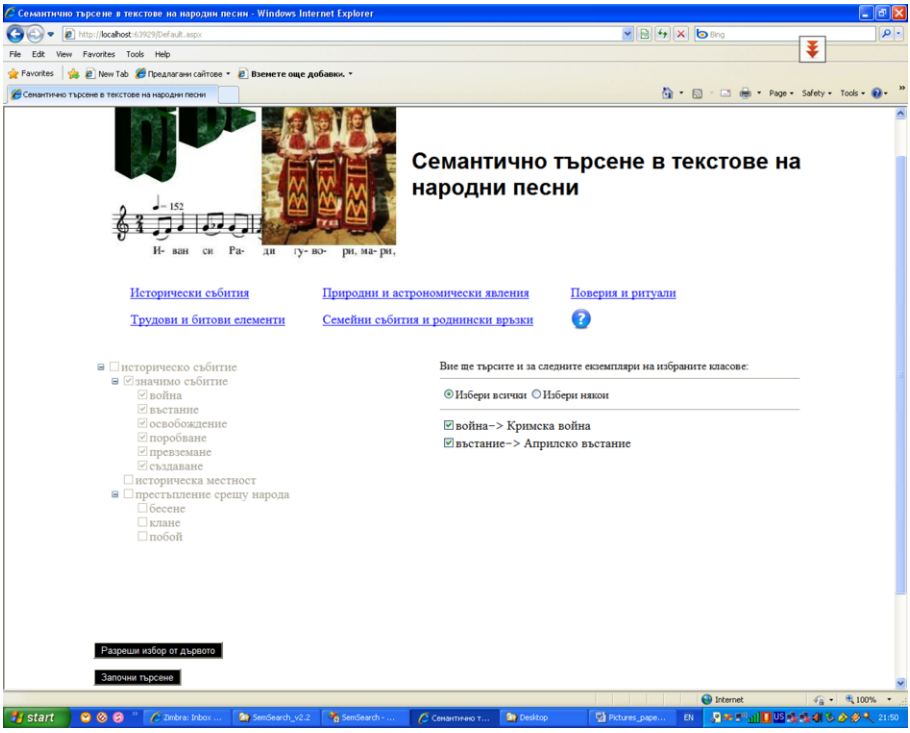

Figure 3. Construction of a user query for ontology-based search (step 2).

5. Implementation

For the development of the subject ontology we used one of the most popular ontology editors – Protégé-OWL [12]. The search engine has been implemented using the following technologies and tools:

- ASP.NET 3.5 technology;
- C# programming language;
- IDE Visual Studio 2008.

The implementation of the basic component of the search engine which realizes semantics oriented (ontology-based) search in the lyrics of songs works on two main classes defined for the pupose:

- the OntologySearcher class searches the ontology for primitive and defined classes that match the original user query as well as for instances of these classes. By doing so, it implements the augmentation of the user query;

- the XMLSearcher class finds the files with catalogue descriptions in a given folder which have "folk_song_text" nodes with values containing at least one word from a given list. It also returns a dictionary of the discovered incorrect catalogue files.

6. An Example

Let us suppose for example that the user defines a query for ontology-based search in the lyrics of songs which concerns the concept "historic event" ("значимо историческо събитие" in Bulgarian). During the execution of this query, first of all it is augmented and refined with the assistance of the user as shown on Figure 2 and Figure 3. Then a consecutive search in the catalogue descriptions of songs follows. As a result, all documents with <folk_song_text> element values containing phrases that are juxtaposed with at least one element of the augmented user query, are extracted. A list with the titles of the discovered songs is properly visualized on the user screen (see Figure 4).

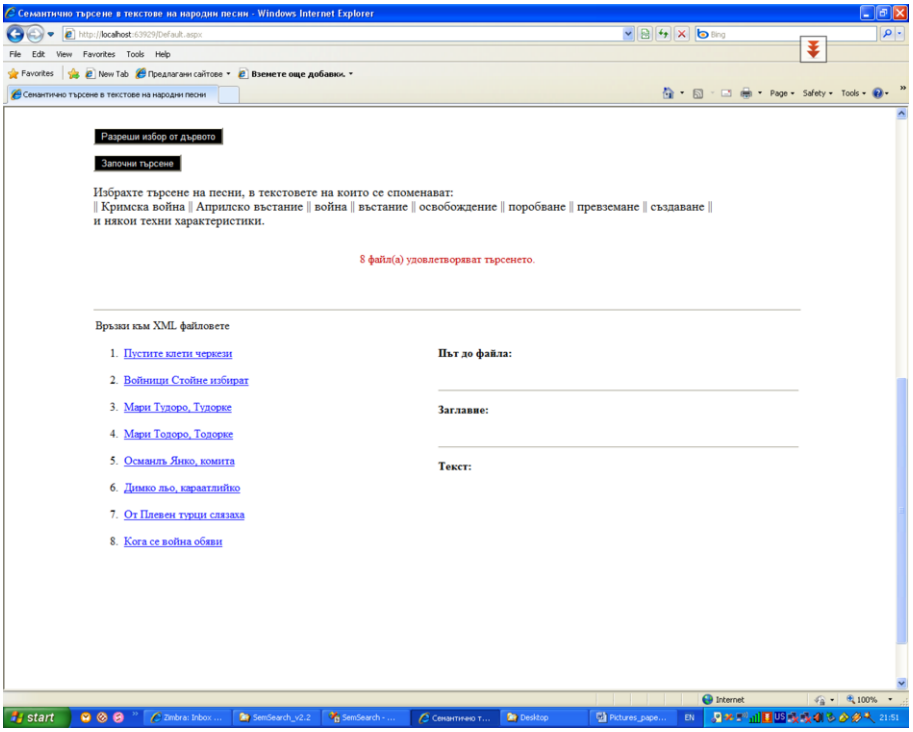

Figure 4. Search results for a user query containing the phrase "historic event" (level 1).

When the user clicks on the name of a particular song satisfying the augmented query, the text of this song is displayed in a new window. The discovered words and

phrases that match (are semantically related to) the query, are highlighted. Figure 5 shows some results for the user query example containing the phrase “historic event”.

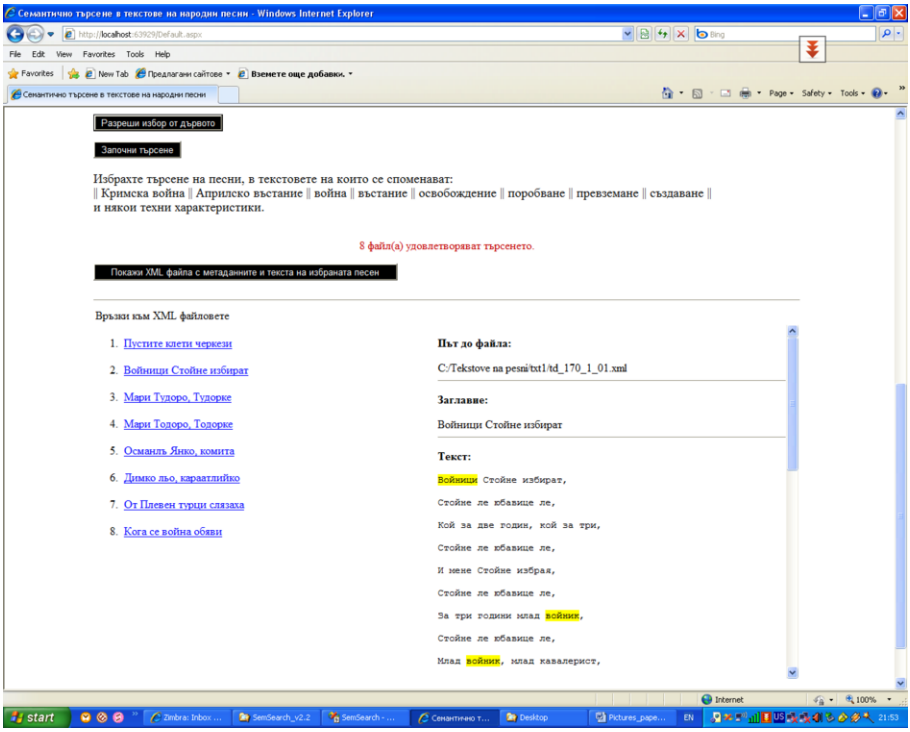

Figure 5. Search results for a user query containing the phrase “historic event” (level 2).

7. Conclusion

The working principles of the search engine of DjDL are designed in order to support its flexibility, interoperability and reusability.

Our current activities are directed to:

- evaluating the performance of the search engine (computing a tentative value of its average precision);
- design and implementation of a tool for flexible and convenient (intuitive) construction of complex user queries without using natural language and proper modification of the search algorithm.

The next step will be to extend the functional facilities of the search engine with a proper tool for semantic search and knowledge discovery in the notes of songs. A main goal in this direction will be to automate the further study of some musical characteristics of Bulgarian folk songs (e.g., their melodies and rhythms) with the aim to discover similarities of songs according to various criteria.

In this way the final version of DjDL will be developed with the aim to provide a complete set of tools which will be useful for a series of further studies in folkloristics, philology and musicology.

Acknowledgements. This work has been supported by the Bulgarian National Science Fund within a project titled "Information technologies for presentation of Bulgarian folk songs with music, notes and text in a digital library", Grant No. DTK 02/54/17.12.2009. The authors are thankful to their student Dicho Shukerov for his contribution to the implementation of the prototype of the search engine of DjDL.

References

[1] R. Guha, R. McCool, E.Miller, Semantic Search. In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*, Budapest, Hungary, 2003, 700-709.

[2] S. Kruk et al., JeromeDL – a Semantic Digital Library. In: J. Golbeck, P. Mika (Eds.), *Proceedings of the Semantic Web Challenge Workshop at ISWC2007*, Busan, South Korea, 2007.

[3] L. Castro, O. Giraldo, A. Castro, Using the Annotation Ontology in Semantic Digital Libraries. In: A. Polleres, H. Chen (Eds.), *Proceedings of the ISWC 2010 Posters & Demonstrations Track*, Shanghai, China, 2010.

[4] S. Kruk et al., The Role of Ontologies in Semantic Digital Libraries. *10th European Conference on Research and Advanced Technology for Digital Libraries*, Alicante, Spain, 2006.

[5] L. Peycheva, N. Kirov, Bulgarian Folk Songs in a Digital Library. In: R. Pavlov, P. Stanchev (Eds.), "*Digital Preservation and Presentation of Cultural and Scientific Heritage. International Conference, Veliko Tarnovo, Bulgaria, 11-14 September, 2011*", ISSN 1314-4006, IMI – BAS, Sofia, 2011, 60-68.

[6] M. Nisheva-Pavlova, P. Pavlov, Semantic Search in a Digital Library with Bulgarian Folk Sons. In: Y. Tonta et al. (Eds.), "*Digital Publishing and Mobile Technologies. 15th International Conference on Electronic Publishing, June 22-24, 2011, Istanbul, Turkey*", ISBN 978-975-491-320-0, Hacettepe University, Ankara, 2011, 103-109.

[7] N. Kirov, Digitization of Bulgarian Folk Songs with Music, Notes and Text. *Review of the National Center for Digitization* **18** (2011), 35-41.

[8] M. Nisheva-Pavlova, Providing and Maintaining Interoperability in Digital Library Systems. In: *Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010)*, ISBN 978-954-07-3168-1, St. Kliment Ohridski Uiversity Press, 2010, 200-208.

[9] M. Nisheva-Pavlova, P. Pavlov, Search Engine in a Class of Academic Digital Libraries. In: T. Hedlund, Y. Tonta (Eds.), "*Publishing in the Networked World: Transforming the Nature of Communication. 14th International Conference on Electronic Publishing, 16-18 June 2010, Helsinki, Finland"*, ISBN 978-952-232-085-8, Edita Prima Ltd, Helsinki, 2010, 45-56.

[10] P. de Juan, C. Iglesias, Improving Searchability of a Music Digital Library with Semantic Web Technologies. In: *Proceedings of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE'2009)*, ISBN 1-891706-24-1, Knowledge Systems Institute Graduate School, Boston, Massachusetts, 2009, 246-251.

[11] L. Peycheva, G. Grigorov, How to Digitalize the Folk Song Archives? *Review of the National Center for Digitization* **18** (2011), 42-58.

[12] H. Knublauch, R. Fergerson, N. Noy, M. Musen, The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. *Third International Semantic Web Conference*, Hiroshima, Japan, 2004, 229-243.