Semantic Search in a Digital Library with Bulgarian Folk Songs

Maria M. Nisheva-Pavlova *and* **Pavel I. Pavlov**

Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., Sofia, Bulgaria.
Email: marian@fmi.uni-sofia.bg, pavlovp@fmi.uni-sofia.bg

Abstract: *The paper presents some aspects of an ongoing project aimed at the development of technologies for digitization of Bulgarian folk music and building a heterogeneous digital library with Bulgarian folk songs presented with their music, notes and text. This digital library will provide both digital preservation of the sound recordings, lyrics and notations of Bulgarian folk songs and a possibility for new interpretations of the archaic Bulgarian folklore heritage. Some facilities of the search engine under development to implement various types of search and access to the library resources are analyzed in the paper. The emphasis of the discussion falls on the tool provided for semantic search in the lyrics of songs.*

Keywords: *Digital library; metadata; semantic web; ontology; search engines.*

Introduction

Bulgarian folk music is a valuable resource of cultural memory and is among the main characteristics of the national identity of Bulgarian people. Throughout the years the Bulgarian researchers of musical folklore have written down hundreds of thousands of musical folk samples (songs and instrumental melodies) in lyrics and notes. Part of these music notations has been published, another part is preserved as manuscripts in specialized institutional or personal archives.

The paper presents some initial results of the activities within an ongoing project aimed at the development of technologies for digitization of Bulgarian folk music and building a digital library (named DjDL) with Bulgarian folk songs presented with their text, notes and music. DjDL will serve as a platform for digital preservation of the sound recordings, lyrics and notations of Bulgarian folk songs which will provide a possibility for exploration and new interpretations of the archaic Bulgarian folklore heritage. The main facilities of the search engine under development to realize various types of search and access to the lyrics of songs are discussed in detail.

Main Characteristics of the Library Resources

Currently DjDL keeps a collection of digital objects which represent a part of the unpublished archive manuscripts of Prof. Todor Dzhidzhev containing recordings of over 1000 folk songs from the Thracia region of Bulgaria. Completion and diversification of this collection with music from various folk musical dialects has been planned for the near future.

After a careful examination, LilyPond[1], extended with tools for using some specific symbols and other elements necessary for notation of the Bulgarian national music (Kirov, 2010), was chosen as a technological base for coding of note records from original manuscripts with field recordings of Bulgarian folk songs.

[1] LilyPond ... music notation for everyone, http://lilypond.org/

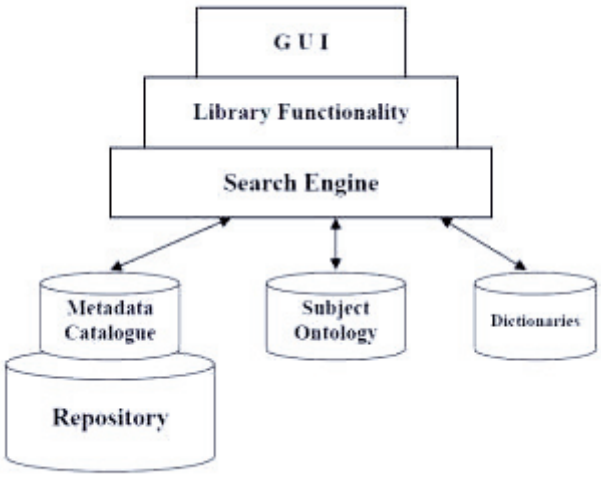

Figure 1. Functional structure of DjDL

DjDL has the typical architecture of an academic digital library with heterogeneous resources. Its functional structure is shown in Figure 1.

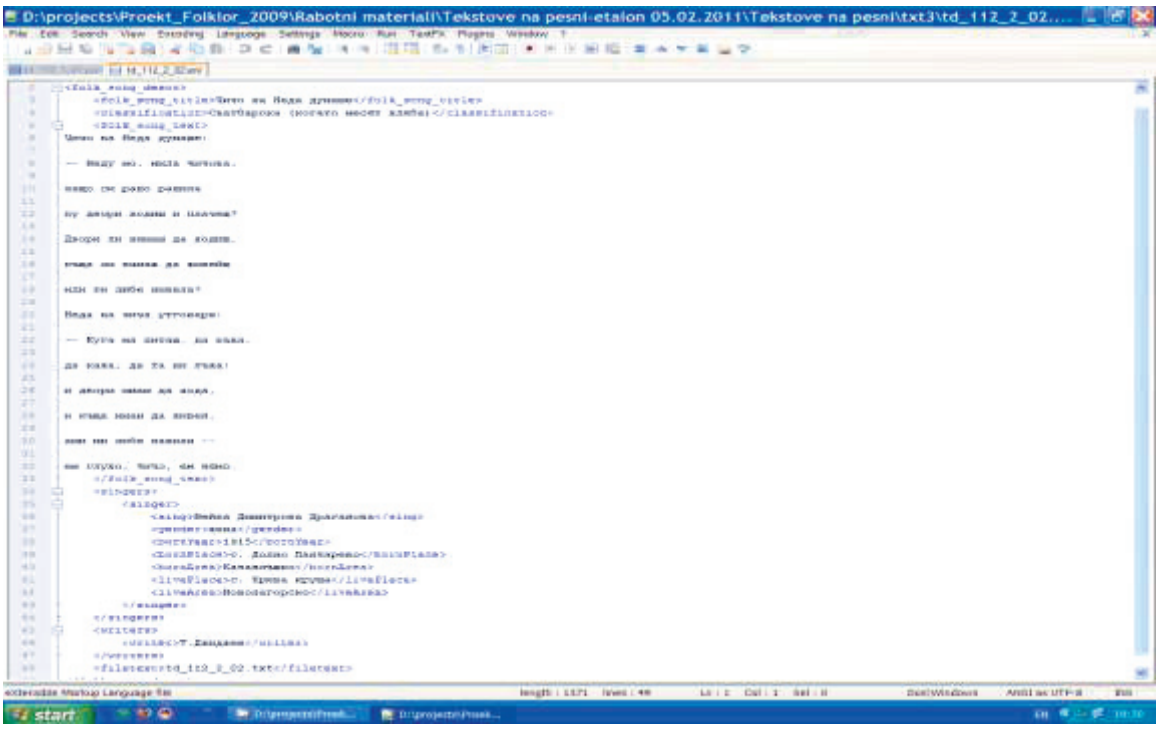

Figure 2. Part of a catalogue description

The library catalogue contains short descriptions (in XML format) of the songs included in the repository. Various types of relevant metadata (metadata attributes) are provided, for example: the title of the song, the song genre according to different classification schemes (e.g. according to the typical time and space of performance, thematic focus(es), context of performance, etc.), the region of folk dialect, data specifying the informant (the person who conveyed the song to folklorists) and the time and place of gathering the song (the folklorist who gathered the song, the singer(s), the time and

place of record, etc.). More precisely, each catalogue entry contains the text (i.e., the lyrics) of a particular song accompanied by the corresponding metadata.

Figure 2 displays a selected part of the catalogue description of the folk song entitled "Чичо на Неда думаше" ("Neda's uncle speaking to her").

The repository of DjDL contains heterogeneous resources of the following types (Peycheva et al., 2010):

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

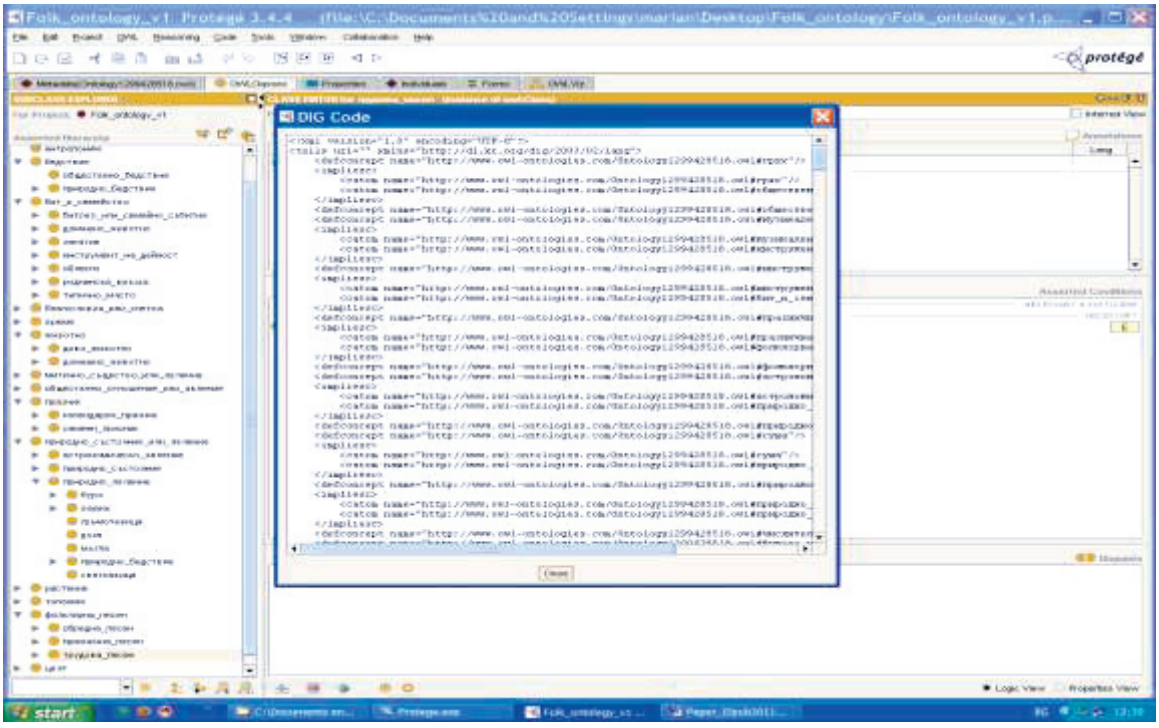

Figure 3. Part of the concept hierarchy and the corresponding DIG code

The subject ontology consists of several interrelated subontologies needed by the search engine of DjDL and developed especially for the purposes of the discussed project:

- 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 mythical creatures and demons;
- ontology of settlements.

For the development of the subject ontology we used one of the most popular ontology editors – Protégé/OWL (Knublauch, 2003; Knublauch et al., 2004). The implementation of the search engine is especially oriented to the corresponding DIG code with two simple automatically made changes in it: reduction of the ontology URL in all class, property and individual names and replacement of the underscore characters with spaces.

Figure 3 shows a part of the "is-a" hierarchy of concepts included in the particular subontologies.

The search engine of DjDL also uses some proper dictionaries available in digital (XML) format – a dictionary of synonyms and a dictionary of obsolete and dialect words.

Functionalities of the Search Engine

The search engine of DjDL supports two main types of search: keywords-based and semantic search. Its current version realizes some facilities for search in the catalogue metadata and the lyrics of songs only. The design and the implementation of this search engine are based on some former results of the authors (Nisheva-Pavlova & Pavlov, 2010) and some ideas from de Juan & Iglesias (2009) and McGuinness (2003). Its full palette of functionalities has been specified after a careful study of the requirements of the typical user groups (specialists and researchers in ethnomusicology, verbal folklore and folkloristics in general, philologists, etc.).

The user queries define restrictions on the values of certain metadata attributes and/or the lyrics of the required folk songs. The search procedure implies the performance of some pattern matching activities in which the catalogue descriptions are examined one by one and those 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.

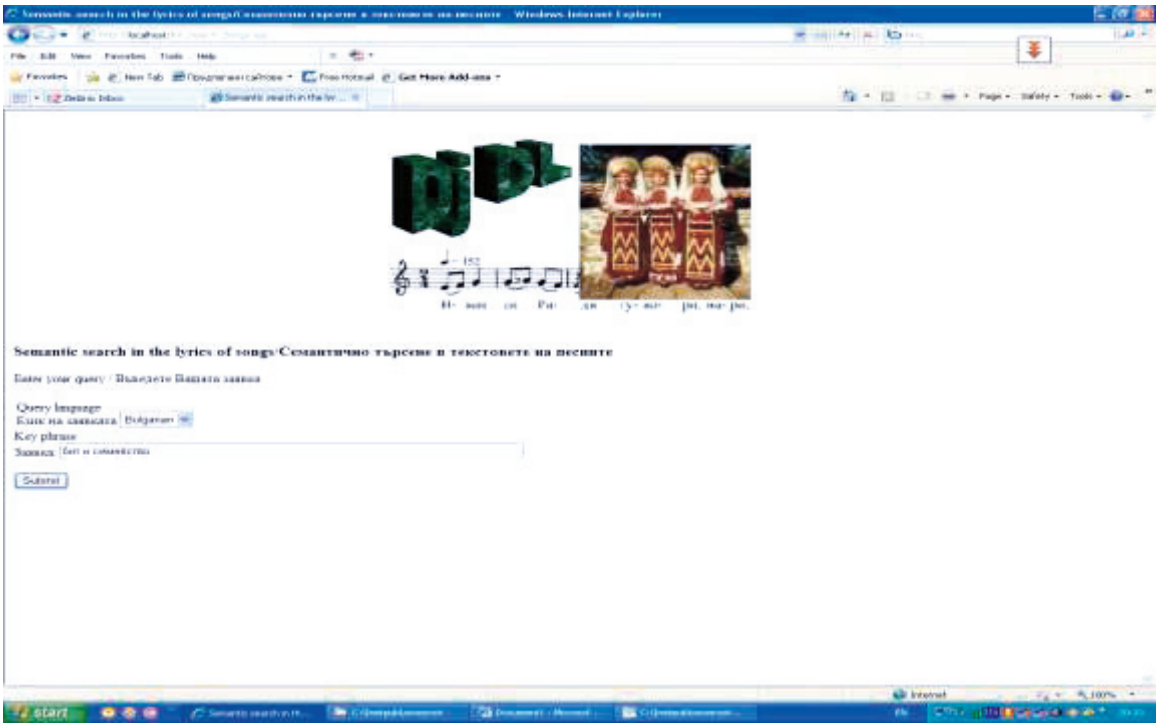

Figure 4. A sample user query for semantic search

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

As a first step in 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 (in the catalogue descriptions) or combined search in the lyrics of songs and catalogue metadata. One may 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 corresponding text. As a result of the user query processing, a list of links to the discovered files with lyrics of songs is properly displayed. This list may be ordered by the titles of songs or by the number of appearances of the words (or phrases) included in the user query.

The following list contains some typical examples of queries for keywords-based search:

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

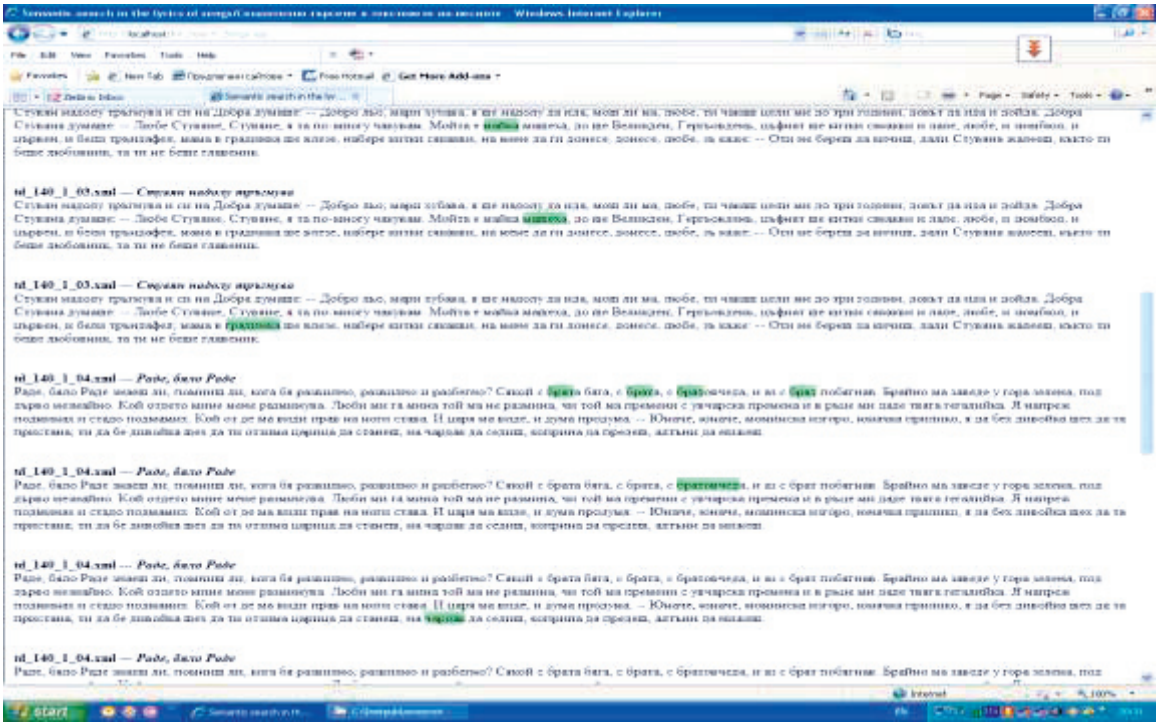

Figure 5. Part of the search results for a user query containing the phrase “popular customs and family”

The semantic search service in DjDL is aimed at the provision of some additional functional facilities for extension, generalization and refinement (automatic reformulation according to the available explicit domain knowledge) of the queries for keywords-based search.

The extension (more precisely, the augmentation) of the user query is based on the use of the subject ontology and the available dictionaries – the dictionary of synonyms and the dictionary of obsolete and dialect words. 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, i.e. the corresponding more specific concepts from the ontology (the concept hierarchy) are added to the one given by the user. The names of the existing instances of these concepts are added as well. Then the search engine adds to the newly constructed set of queries some synonyms, derivatives, dialect and obsolete forms of the main terms found in the mentioned dictionaries. Thus the user query is augmented as far as possible and in fact has the shape of a disjunction of all included forms of concepts and instance names. In this form it is ready for further processing.

Figure 4 illustrates a sample user query for semantic search containing the phrase “popular customs and family” (“бит и семейство” in Bulgarian).

Figure 5 shows a screenshot displaying part of the search results for this query which contains the titles of some retrieved songs (the text in bold italic type) and their texts in a compressed form as well as the corresponding file names (the names of the corresponding catalogue entries – the text in bold face). The discovered words and phrases that are semantically related to the original user query are highlighted.

A careful examination of the search results displayed in Figure 5 shows that some of the highlighted words denote concepts which are more particular than the query one (in accordance with the "is-a" hierarchy described within the subject ontology) – typical scenes of narrative in folk songs („чардак") and ties of relationship („майка", „мащеха", „брат", „братовчед"). The word „градинка" is a diminutive form of „градина" ("garden") which is also a typical scene of narrative.

As examples of queries for semantic search of interest for folklorists (according to Peycheva & Grigorov, 2010) that can be executed by the search engine of DjDL, one may indicate the queries for search and retrieval of:

- songs devoted to important historical events or 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 important family events or typical family relationships (daughter-in-law – mother-in-law, son-in-law, mother-in-law, etc.) or joyful/unfortunate family events are mentioned.

The current implementation of the search engine also provides some facilities for processing of user queries containing 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 a particular singer (grouped by the names of singers);
- songs performed by singers born in a particular settlement or region (grouped by settlements/regions);
- songs performed by singers who have moved from a particular settlement or region;
- songs performed to the west/east/north/south of a specific settlement/region;
- songs performed in a specific region (grouped by regions of performance);
- songs in which at least/more than/exactly a specific number of toponyms/anthroponyms are mentioned.

In contrast with our former experience in implementation of similar techniques for the purposes of semantic search in collections of digitized mediaeval manuscripts (Pavlov & Nisheva-Pavlova, 2006) and archival documents (Nisheva-Pavlova et al., 2007), the results obtained in the discussed case of search in the texts of folk songs may be evaluated as insufficient. We suppose that the reason for this conclusion lies in the vast use of similes, metaphors, idioms and other sophisticated or language-dependent stylistic devices in the folklore lyrics. With the aim of overcoming some aspects of this problem, we intend to design and use a set of proper patterns of typical stylistic or thematic constructs which could be matched with relatively large parts of the texts of folklore songs. For example, we have already defined a number of patterns of constructs standing for "unfaithfulness", "jealousy", "discontent" and "sedition" as well as the corresponding pattern matching rules and recently performed various experiments with them.

Our current activities are directed to the design and implementation of a tool for flexible and convenient (intuitive) construction of complex user queries (without using natural language), in particular queries including proper combinations of the example types mentioned above.

The next step will be to extend the functional facilities of the search engine of DjDL 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 of discovering similarities of songs according to various criteria.

Summary and Future Work

The final version of DjDL will be developed with the aim of providing a complete set of tools which will be useful for a series of further studies in folkloristics, philology and musicology, for example:

- creation of a set of the phonetic variations which every sound can undergo in accordance with the Bulgarian lexicology and dialectology;
- classification of folk songs according to different criteria: technical, implicit-musical, implicit-textual, statistical, genre, cultural, etc.;

- creation of frequency dictionary and concordance, based on the lyrics of songs and inferring of theoretical conclusions from the obtained results.

Acknowledgements

The project "*Information technologies for presentation of Bulgarian folk songs with music, notes and text in a digital library*" has been supported by the Bulgarian National Science Fund under Grant DTK 02/54.

References

De Juan, P. & Iglesias, C. (2009). Improving searchability of a music digital library with semantic web technologies. *Proceedings of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE'2009)* (pp. 246-251). Boston, Massachusetts: Knowledge Systems Institute Graduate School, ISBN 1-891706-24-1.

Kirov, N. (2010). Digitization of Bulgarian folk songs with music, notes and text. To appear in: *Review of the National Center for Digitization*, ISSN 1820-0109. Retrieved, April 3, 2011, from http://eprints.nbu.bg/627/1/NKirov_folk.pdf.

Knublauch, H. (2003). An AI tool for the real world: Knowledge modeling with Protégé. *JavaWorld*, June 20.

Knublauch, H., Fergerson, R., Noy, N., & Musen, M. (2004). The Protégé OWL plugin: An open development environment for semantic web applications. In S.A. McIlraith et al. (Eds.). The Semantic Web: ISWC 2004. *Lecture Notes in Computer Science*, 2004, Volume 3298/2004, 229-243. Heidelberg: Springer.

McGuinness, D. (2003). Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman, & W. Wahlster (Eds.), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Cambridge, MA: MIT Press, ISBN 978-0-262-06232-9.

Nisheva-Pavlova, M., Pavlov, P., Markov, N., & Nedeva, M. (2007). Digitisation and access to archival collections: A case study of the Sofia Municipal Government (1878 – 1879). L. Chan and B. Marten (Eds.) *ELPUB2007. Openness in Digital Publishing: Awareness, Discovery and Access - Proceedings of the 11th International Conference on Electronic Publishing held in Vienna, Austria 13-15 June 2007* (pp. 277-284). Vienna: ÖKK-Editions, ISBN 978-3-85437-292-9.

Nisheva-Pavlova, M. & Pavlov, P. (2010). 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* (pp. 45-56). Helsinki, Edita Prima Ltd, ISBN 978-952-232-085-8.

Pavlov, P. & Nisheva-Pavlova, M. (2006). Knowledge-based search in collections of digitized manuscripts: First results. In Bob Martens and Milena Dobreva (eds.). *Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing. Bansko, Bulgaria June 14-16, 2006* (pp. 27-35), Sofia: FOI-Commerce, ISBN 978-954-16-0040-5.

Peycheva, L. & Grigorov, G. (2010). How to digitalize folklore song archives? To appear in: *Review of the National Center for Digitization*, ISSN 1820-0109.

Peycheva, L., Kirov, N., & Nisheva-Pavlova, M. (2010). Information technologies for presentation of Bulgarian folk songs with music, notes and text in a digital library. *Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (ISGT'2010)* (pp. 218-224). Sofia: St. Kliment Ohridski University Press, ISBN 978-954-07-3168-1.