DIGITAL LIBRARY FOR BULGARIAN TRADITIONAL CULTURE AND FOLKLORE

D. Paneva-Marinova [a], R. Pavlov [a], K. Rangochev [a]

[a] Institute of Mathematics and Informatics—Bulgarian Academy of Sciences, Bl. 8, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria—(dessi, radko)@cc.bas.bg, krangochev@yahoo.com

KEY WORDS: Bulgarian Folklore, Multimedia Digital Libraries, Functionality, Services

ABSTRACT:

The Bulgarian folklore digital library is built during the "Development of Digital Libraries and Information Portal with Virtual Exposition 'Bulgarian Folklore Heritage'" module of the national research project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage". This Internet-based environment is a place where folklore objects (mainly from the Funds of the Institute for Folklore at the Bulgarian Academy of Sciences) of different kinds and origins were documented, classified, and „exhibited" in order to be widely accessible to both professional researchers and the wide audience. This paper describes the library, its main services and components, their implementation and testing procedures. The paper also discusses problems arising during the formalization of folklore knowledge and the Bulgarian folklore ontology development, needed for semantic annotation of folklore objects in the library.

1. INTRODUCTION

Preserving and presenting the national folklore heritage is a long-term commitment of scholars and researchers working in many areas. From centuries every generation aims to keep records about work and social life, so that they could be revised and studied by the next generations. For a long time this heritage has been maintained in libraries, museums and research laboratories, where not everyone was able to access this wealth.

New information and multimedia technologies that have been developed during the past couple of years introduced new methods of documentation, maintenance and distribution of the huge amounts of collected material. Among these new technologies are digital libraries, which have already proven their worth as a contemporary conceptual solution for access to information archives. Digital libraries contain diverse hypertext-organized collections of information (digital objects such as text, images, and media objects) that are organized thematically and are managed by complex specialized services such as content structuring, advanced search (semantic-based search, multilayer and personalized search, context-based search), resources and collection management, information retrieval, indexing, semantic annotation of digital resources and collections, content grouping, metadata management, personalization and adaptive access, multilinguality, digital information protection and preservation, tracking services, etc. (Pavlov et al., 2006). Digital libraries enable "any citizen to access all human knowledge any time and anywhere, in a friendly, multi-modal, efficient, and affective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices" (Brainstorming report, 2001).

In an attempt to answer the need of wider accessibility and popularization of Bulgarian folklore culture, a team from the Institute of Mathematics and Informatics has developed a Bulgarian folklore digital library (BFDL) within the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" national research project (Folknow). The project aims to develop a complete web-based environment for a virtual presentation of the Bulgarian folklore treasure kept in the funds of the Institute for Folklore of the Bulgarian Academy of Sciences.

This paper presents the Folknow project, its vision and ideas. The functional specification, implementation and testing of the Bulgarian folklore digital library are described, extending the presentation made in (Rangochev et al., 2007). The paper also discusses problems arising during the formalization of folklore knowledge and the Bulgarian folklore ontology development, needed for semantic annotation of folklore objects in the library.

2. FOLKNOW PROJECT

The project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" started 4 years ago with fundamental research on contemporary technologies for virtual exposition of intangible cultural heritage. Its aim is to build an experimental release multimedia digital library for the registration, documentation, and access to a wide range of Bulgarian folklore objects. The complex structure and the multi-layer characters of the folklore objects require an innovative approach for knowledge representation. The rich-in-content web-presenting of the Bulgarian folklore knowledge defines the usage of modern methods and technologies for developing a digital archive, which will be used not only for preservation and access to the information, but as a tool for scientific research analysis development. The main project's tasks are to create a digital library and information artery with semantic-sensible inferring maintenance in order to present in virtual form the valuable phenomena of the Bulgarian folklore heritage in a comprehensive and easy-to-use way. The realization of the project gives a possibility for wide social applications of the multimedia collections for the purposes of interactive distance learning/self-learning, research activities in

the field of Bulgarian traditional culture, and for cultural tourism and ethnotourism in Bulgaria.

3. COMPLEXITY OF THE BULGARIAN FOLKLORE DOMAIN

The folklore knowledge and therefore the ethnological research have a systemic character (Rangochev, 1997). Since the early period of Bulgarian ethnology until the present day, scholars describe, investigate, and analyse different descriptive schemas for this knowledge. As a rule, scholars study a certain area of knowledge in a particular topos of the Bulgarian ethnical territory and find out an algorithm (where there is a process) or a structural description and afterwards the procedure is repeated in another topos, etc. Finally, a summarized algorithm or a structural description is achieved which is – as a matter of fact – the research abstraction (for instance, the "full" description of the "Bulgarian koleduvane" (Christmas rites) is an algorithm which does not coincide with its local variants). All this means that the ethnological studies are hierarchically organized. Leaving the particular topos (a village, for instance), the scholar focuses on bigger entities (such as a region, ethnographical region, or an ethnical territory) and thus he deals with a model of the studied area of ethnological knowledge. A danger in this hierarchical modelling could be the possibility to neglect important systematic links of knowledge (For instance, if we consider some folklore paradigms of kinship, it can turn out that the same person is involved in several systems of kinship: 1. by blood: grandson- son-brother-uncle; 2. by rite: brother-in-law; 3. by profession).

Another problem comes from the specifics of fieldwork investigations. As a rule, the scholar extracts parts of the ethnological knowledge by the means of interview with the informants. Therefore, ethnologists study phenomena which are not person-specific but characterize the community but they use for this purpose the memories and opinions of particular people.

Another important problem is the specifics of the ethnological research: these types of studies are mostly abstract, due to several historical, objective and subjective reasons (technology of recording, ethical, ideological, and scholarly prejudices, etc.). The records of samples of Bulgarian folklore which are studied by scholars in practice contain partial information: for instance, songs have been recorded only as texts without notation; or there is no information for the discourse practices conveying the oral narratives; or in many cases the records are made by means of structured interviews and not by inclusive interviews. Therefore, the conclusions of scholars are usually based on partial information (Rangochev, 1997).

All these problems require new flexible methods for representation of knowledge in formal and single structures for securing manners of access and management of this knowledge. In order to formally represent the folklore knowledge the ontology of the Bulgarian folklore was produced.

4. ONTOLOGICAL PRESENTATION OF THE FOLKLORE KNOWLEDGE

Originally, the term ontology comes from philosophy where it is employed to describe the existence of beings in the world. In 1993, Gruber's definition becomes the most referenced on the knowledge technologies literature: "ontology is a formal, explicit specification of a shared conceptualization". Conceptualization refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. Explicit means that the type of concepts used and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable, which excludes natural language. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Ontologies can be used for many different purposes. The literature on knowledge representation contains research on the use of ontologies for data-interchange, for data-integration, for data-querying or for data visualization. In general, visualization of information can be seen as a two-step process. In a first step, information is transformed into some intermediate semantic structure. This structure organizes the raw information into a meaningful structure. In a second step, this semantic structure is used as the basis for a formal visual representation. We used this approach in our work on the Bulgarian folklore ontology development.

To efficiently represent the folklore annotation framework and to integrate all the existing data representations into a standardized data specification, the folklore ontology need to be represented in a format (language) that not enforce semantic constraints on folklore data, but can also facilitate reasoning tasks on folklore data using semantic query algebra. This motivates the representation of Bulgarian folklore ontological model in Web Ontology Language (OWL). OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with a formal semantics. Knowledge captured from folklore data using OWL is classified in a rich hierarchy of concepts and their inter-relationships. OWL is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying. We investigated the use of OWL for making Bulgarian folklore ontology using Protégé OWL Plug-in.

4.1 Ontology of the Bulgarian Folklore

Since one of the targets of the Folknow project is to present the valuable phenomena of the Bulgarian folklore in suitable virtual form using knowledge technologies, we have to observe and specify the experience that has been gained in the last 500 years in the area of traditional folklore i.e. to construct Bulgarian folklore domain ontology.

FolkKnow annotator/indexers using this ontology will semantically describe and index the raw audiovisual content in order to create and maintain reusable digital objects for the BFDL.

The ontology will be also used to realize semantic-based access to concrete digital objects, representing folklore objects, described by their main features, technical data or context. All this information is included within the Folklore Ontology Concept – the root concept for the ontology.

The process of building of the Bulgarian folklore ontology for the Folknow project is necessarily iterative. The first activity is the definition of the scope of the ontology. Scoping has been mainly based on several brainstorming sessions with folklorists and content providers. Having these brainstorming sessions allowed the production of most of the potentially relevant terms.

At this stage, the terms alone represented the concept, thus concealing significant ambiguities and differences of opinion.
A clear issue that arose during these sessions was the difficulty in discovering of definite number of concepts and relations between these concepts. The concepts listed during the brainstorming sessions were grouped in areas of work corresponding naturally arising sub-groups. Most of the important concepts and many terms were identified. The main work of building the ontology was then to produce accurate definitions.

The folklore object is related to two levels of knowledge, enriched with a set of sub-levels of the data classification. All these levels of knowledge or "thematic entities" in the ontology conception are supported by the scientific diagnosis results and the related documentation. The entity "Identification and description" consists of general historical data, identifying aspects such as title, language, archival signature, period, current location of the folklore object, annotation, first level description, second level description, etc. The entity "Technical" includes technical information both revealing the technologies used for folklore object capturing and recording, record situation, record type, record place, record date, main participants in the process (record maker and informant), etc. Figure 1 depicts part of the main concepts and properties in the Bulgarian folklore ontological model.

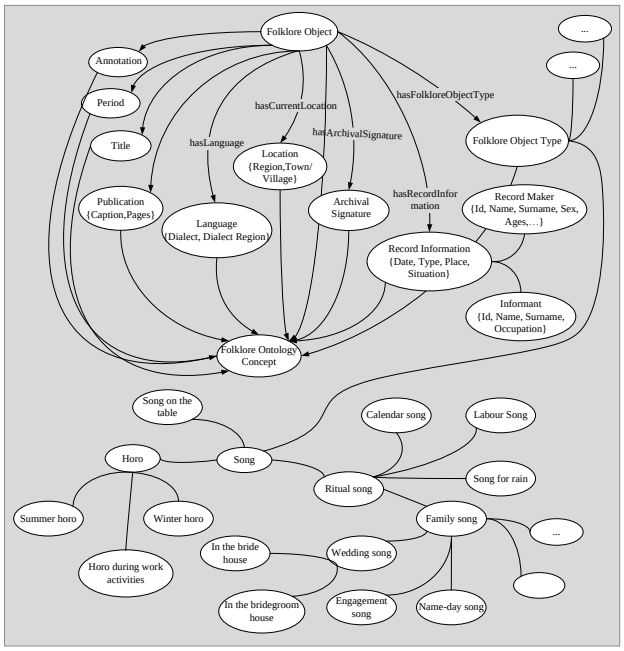

Figure 1: Part of the main concepts and properties in the ontology of the Bulgarian folklore

A detailed description of the ontology of the Bulgarian folklore is made in (Paneva et al., 2007; Luchev et al., 2008).

4.2 Semantics of a Complex Folklore Object

According to the classificatory categories in the Funds of the Institute of folklore of the Bulgarian Academy of Sciences the folklore objects are simple (for example, "song", "food", "magic", etc.) and complex (archived objects which cannot be so clearly and unanimously classified: the same units (parts), according to its informational content, could be classified into different Folklore Object Types).

Example of a complex folklore object is CFO A1_146_2-14, an interview containing information of the catholic community in the village of Oresh, Svishtov region, northern Bulgaria (see figure 2). The emphasis in the interview is on the ritual, festival, and everyday life in the village, on the popular beliefs and knowledge. Every one of these folklore object types has several sub-categories, depicted on figure 2.

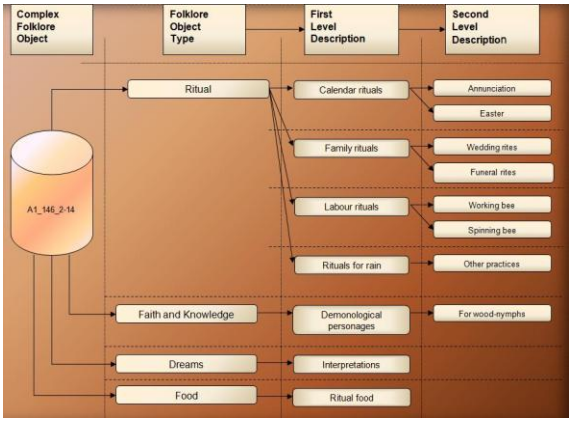

Figure 2: Example of a complex folklore object

The complex folklore object A1_146_2-14 annotated according to the ontology of the Bulgarian folklore has four subclasses of the Folklore_Object_Type class, in particular: "Ritual", "Faith and Knowledge", "Dreams" and "Food". They have also the following sub-subclasses: "Calendar Rituals", "Family Rituals", "Labour Rituals", and "Rituals for Rain" (Ritual class); "Demonological Personages" (Faith and Knowledge class); "Interpretations" (Dreams class); "Ritual" (Food class). On the next level, this complex folklore object is semantically represented by "Annunciation", "Easter", "Wedding rites", "Funeral rites", "Working Bee", "Spinning Bee", and "Other practices" in Ritual sub-subclass; further it is represented by "Wood-nymphs" in Faith and Knowledge sub-subclass, etc.

5. FUNCTIONAL SPECIFICATION OF THE BULGARIAN FOLKLORE DIGITAL LIBRARY

The key for the current release of BFDL is the efficiency and the provision of strictly designed functionalities, powered by a long-term observation of the users' preferences, cognitive goals, and needs, aiming to find an optimal functionality solution for the end users. In BFDL we also follow the requirements of experts in the area of Bulgarian folklore and the accepted functional specification for a digital library. Following them the basic BFDL functional modules are:

- A module for adding and editing folklore objects. The library expects as an input two types of objects: simple folklore objects and complex folklore objects.
- A module for viewing the content of folklore objects (according to their base type and rubric to which they belong or by different descriptive characteristics). Figure 3 shows a snapshot of a folklore object.
- A module for searching by: signature and archive number; keywords of the following categories: name, language, annotation, type of the folklore object/rubric; file type;

record information (simultaneously or one by one): by situation, by reporter name, by recorder name, by record date and by recording location; extended search – it provides the option for searching through all the object characteristics;

Figure 3: Folklore object preview

- A module for managing the user data;
- A module for monitoring the user's actions, which keeps track of the following: a) Actions related to working with the system: registration, logging in the system, unsuccessful log-in attempts, logging out, changing of the user password, e-mail address change, etc.; b) Actions related to the object manipulation: adding an object, editing an object, deleting an object, adding a file, deleting a file; c) Actions related to the content viewing: review of objects by their characteristics, view of a single object, searching for objects by characteristics; d) Other administrative actions: changing the user's level, deleting a user, generation of an XML copy of the data in the system;
- A module for file format conversion;
- A module for generation of XML copies of the objects in the system.

The module for viewing the content of folklore objects is available to all users of the library, except the administrators. The reason is that the administrators of such systems are often people who don't have any relation to their content; they only do support tasks. The module itself was implemented similar to the Windows OS file browser and KDE, so that it is closer to the familiar user interfaces for viewing hierarchical information. The left side shows a tree of all classes, which inherit "Type of folklore object", and the right side shows a list of objects of the selected class in the tree.

The module for creating and editing folklore objects is used for adding new objects and modifying the information of already created objects. Through it, one can add more multimedia files to an object or delete existing ones.

Searching for information is the most frequent search and therefore the most important operation in a digital library. This is why there are several modules for searching by different criteria:

- Searching by a signature or archive number – This search module is useful for finding objects by their archive number (for example, AIF No 200, folder 1, page 57). In general, there is only one search result. In case of incorrect data, it is possible to have several objects as a result.
- Search by a keyword in the object properties – by name, language, annotation and type of the folklore object – Searching is performed simultaneously over all these properties. It is expected that this module is the most frequently used one. This is why special attention has been paid to its optimization.
- Searching by record information – This module is used to find all the objects which cover some of the following conditions: all the objects recorded in a given situation, for example an interview, chat/conversation, etc.; all the objects recorded by a given person; all the objects recorded by a given informer; all the objects recorded in a given period of time; all the objects recorded in a given location.
- Searching by file type – This module allows getting a list of all the objects to which there is a multimedia file attached – audio, video or images. This type of searching uses the database in which the administrative information is stored instead of the OWL file that contains the ontology.
- Complex search on all fields semantically describing the folklore object. Using this search simple and complex folklore objects could be found, tracking their semantic metadata records.

Most types of searching use SPARQL (SPARQL Protocol and RDF Query Language). This is a language for requests to the RDF and OWL ontologies. The language is in a standardization process by RDF Data Access Working Group as an official recommendation of the World Wide Web Consortium. The SPARQL syntax is similar to the most widespread language for database requests – SQL.

The module for monitoring the user's actions is intended to keep logs of the objects modified and deleted by the users, so that in case of data deleted by mistake or entered wrongly, the responsible user can be found. There is also a log of search requests, whose purpose is to enable statistical reports about the search types that are used least and most often. It would allow the removal of the rarely used search types and the priority optimization of the ones that are used most often.

The module for file format conversion was developed to provide the ability to present every file which is unsuitable for internet preview in a "light" and convenient form for web preview. The module recognizes the "inconvenient" files, tries to covert them and on success replaces the original file with the new "lighter" file; on failure, the module keeps the original file in the library. The module for generating an XML copy of the data is available only to the system administrators. The purpose behind this module is creating a copy, which can be used as an archive copy on one hand and on the other hand it may serve as raw data for other systems using information from the library.

The presented BFDL functionality aims to serve a wide range of users – specialist and non-specialist. The group of specialists is composed by scientists who study Bulgarian folklore professionally and search for specialized information on the observed folklore objects. The group of non-specialists has interests and wants only to learn more about the classical Bulgarian folklore objects. The BFDL system supports several users' levels: administrators, folklore content editors, specialist viewers and non-specialists viewers. Their individual characteristics, needs, interests, motivation, and preferences are discussed in (Pavlov et al., 2006).

6. IMPLEMENTATION AND TESTING THE BULGARIAN FOLKLORE DIGITAL LIBRARY

A module for adding objects to the BFDL – Adding objects is implemented through filling and sending a form to the web server. Because of the great number of fields to fill, the form is not generated completely. Only the fields necessary for the creation of the objects are generated, following the semantic descriptions presented in the BFO, built at the first stage of module 3 of the project. The technology used for the implementation is AJAX. The user interface passes a request to the server, in which it requires only that part of the form which according to the user is necessary to create the object. The server processes the request and returns the required fields as a result, which is visualized in the user interface. After all the fields are filled, the user submits the form. The server validates the data and if everything is correct, it adds the object to the data storage. If there is something wrong, it returns a message to the user, relative with the error (usually, an empty field or unacceptable field value). After the server adds the information from the form to the data storage, there follows a check for attached files in the user request. If there are attached files, the server checks if there are file formats which are unsuitable for web presentation (for example, wav, .doc, .mpg, .avi, .mpeg, etc.) and if it finds such files, the system refers to the module for file format conversion to formats suitable for web preview. For each of these files, the module for file format conversion tries to convert them. Upon success, it adds the converted file to the library. On failure (which can occur if the added file has any specifics which the system cannot recognize), it adds the original file to the library. At the end of the object adding procedure, the system refers to the module for monitoring the user actions, where it adds an "object added" event and records the author (the user who created the object) and the event date.

A module for editing objects in a BFDL – The module for editing objects works almost in the same way as the module for adding objects. The difference is that the system doesn't add information about a new object, but replaces the existing information about an object with the new information, provided by the module for editing. Again, the system checks the form for errors, processes the files (if there are new files added), changes its data and finally adds an event for modified object through the module for monitoring the user's activity.

A module for viewing the content of folklore objects – This module takes a request from a user, in which the user specifies the property by which a folklore objects must be found. The module refers to the data storage and makes a request for selecting and sorting the objects by this property. The module for monitoring the users' actions records the "view objects by" event and adds data about the date, the user and the property by which objects are listed. The storage processes the request and returns a result, which the system processes and sends to the user. The user interface visualizes the result in a proper manner.

A module for searching – This module allows the user to set a property or properties by which objects are found. The following algorithms are used:

The algorithm for searching by a single property – The user interface sends a request to the data server specifying the property and its needed value. The module for searching refers to the data storage of semantic metadata with a query for selection and sorting the objects with the needed value of the specified property. The module for monitoring the user actions records the "search" event with the provided search parameters, the date and the user, who performs the search. The storage processes the request and returns a result, which is then processed by the search module and displayed in a proper manner by the user interface.

The algorithm for searching by more than a single property – The algorithm is parallel to the one described above, with the only difference that the query to the data storage is more complicated – there are multiple selections of objects for each search property and the result is a sorted section of the selection results.

After an analysis of the means and standards in the technological implementation of the library environment and the functional modules, the following software was chosen: Operating system: Microsoft Windows Server 2008 x64 Standard; Web server: Apache HTTP Server v 2.2, PHP v 2.2.9; Database management system: MySQL v 5.1 Standard; Tools for the additional modules: FFMPEG, vwWare, HTML, JavaScript, AJAX; Database query language: SPARQL.

The functional components of the architecture of the BFDL were implemented and tested for errors and speed on a server platform with the following hardware configuration: CPU: 2 x Intel QuadCore 2.8 GHz; RAM: 8GB DDR3; HDD: 4 x 500GB, RAID 10 SATA II; LAN: 2 x 1000Mbit.

Testing the functional module for adding/editing a folklore object – Server response time (average of 50 attempts): 0.0058 s, i.e. in theory the functional module for adding/editing an object can process about 172 requests per second for each processor core, which makes 172*8=1376 requests.

Testing the module for viewing folklore objects – Time for server response: 0.009 seconds per request, i.e. 888 requests per second.

Testing the module for searching by a single property – Time for server response: 0.008 seconds per query, i.e. 1000 requests per second.

Testing the module for searching by several properties – The test was performed with 25 properties (it will happen very rarely). Time for server response: 0.01 seconds per query, i.e. 800 requests per second.

Testing the module for file format conversion – Converting video files: the server sends a response before it converts the video file, because the process is relatively slow. The average time of processing a video file is about 30 seconds, i.e. you can add about 16 video objects per minute. In this way, after adding

a video object, its actual recording in the BFDL happens in 30 seconds.

Converting audio files – The server responds before the file is actually processed. The average time for processing an audio file is about 10 seconds, i.e. in theory a system with such a configuration can process about 48 audio files per minute.

Converting MS Word (.doc) files – The conversion takes place in real time. The average server response time is 0.04 seconds per request, which are about 200 requests per second.

7. REFERENCES

Digital Libraries: The Future Directions for European Research Programme, *Brainstorming Report*, San Casioano, Italy, 2001.

Luchev, D., Paneva, D., Rangochev, K., 2008. Use of Knowledge Technologies for Presentation of Bulgarian Folklore Heritage Semantics. *Information Technologies and Knowledge*, 2(4), pp. 307-313.

Paneva, D., Rangochev, K., Luchev, D., 2007. Knowledge Technologies for Description of the Semantics of the Bulgarian Folklore Heritage. In: *Proceedings of the Fifth International Conference „Information Research and Applications" – i.Tech 2007*, Varna, Bulgaria, vol. 1, pp. 19-25.

Pavlov, R., Paneva, D., 2006. Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, *Cybernetics and Information Technologies*, 6(3), pp. 51-62.

Rangochev, K., 1997. Structural Specifications of the Folklore Historical Knowledge. *Problems of the Bulgarian folklore*, Sofia, vol. 10, pp. 444-449.

Rangochev, K., Paneva, D., Luchev, D., 2007. Bulgarian Folklore Digital Library, In: *Proceedings of the Jubilee International Conference on Mathematical and Computational Linguistics "30 years Department of Mathematical Linguistics"*, 2007, Sofia, Bulgaria, pp.119-124.

8. ACKNOWLEDGEMENTS

This work is supported by the NSF of the Bulgarian Ministry of Education and Science under grant No IO-03-03/2006 "Development of Digital Libraries and Information Portal with Virtual Exposition 'Bulgarian Folklore Heritage'" from the project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage".