Abstract—The paper discusses experience of building knowledge portals providing the systematization and integration of knowledge and information resources on humanities, as well as the content-based access to them. To provide a sufficiently complete and consistent representation of the knowledge and information resources, their systematization and integration are performed on the basis of ontology. The suggested approach to building knowledge portals based on ontology was successfully used in the projects aimed at development of the scientific knowledge Internet portals on archeology and computational linguistics.

Index Terms—knowledge portal, ontology, information resources on humanities, content-based access.

I. INTRODUCTION

A steady volume increase of information in various areas of knowledge makes the problem of efficient access to knowledge and information resources more and more actual. Meanwhile, knowledge and information resources on humanities are mainly presented as text documents (in the electronic archive of research organizations) or information resources (in specialized Internet catalogues or portals) and sometimes they are simply allocated on independent humanitarian and technical sites. This fact complicates the search and use of such knowledge and information resources. At the same time, researchers often need an efficient access to scientific papers and other information resources containing description of methods and approaches developed in the framework of a particular humanitarian discipline interesting to them.

To meet the requirements described above, we have suggested the concept and architecture of a specialized knowledge portal [1] based on ontology [2], [3] which is a consistent system of concepts accepted in a certain field of knowledge. Such portals allow one to provide systematization and integration of knowledge and information resources related to the modeled field of science, as well as the content-based access to them.

The paper presents the technology of building knowledge portals and the experience of its application.

This work was supported in part by the Presidium of the Russian Academy of Sciences under Integration Grant SB RAS No. 2/12 and by the Russian Foundation for Basic Research under Grant No. 09–07-00400.

Yu. A. Zagorulko is with the A.P. Ershov Institute of Informatics Systems, Russian Academy of Science, Novosibirsk, Russia (telephone: 383-332-8359, e-mail: zagor@iis.nsk.su).

II. INFORMATION MODEL OF KNOWLEDGE PORTAL

To support a unified representation of heterogeneous knowledge and data and their connectivity, as well as to provide a required functionality of knowledge portals, the information model was suggested. This model joins the subject domain model with the problem domain model of the knowledge portal and describes the types of information presented in its content. Based on the information model, the portal internal data base is built and filling of the portal content, navigation through and search in the portal information space are organized.

Formally, the information model \( M_P \) is described as \( M_P = \{ O_P, MIC_P \} \), where \( O_P \) is the portal ontology, \( MIC_P \) is the model of the portal content.

The ontology \( O_P \) is a core, the basic component of the information model of the portal. It is not only describes the portal knowledge system, but defines the formal structure for representation of its content.

A formalism suggested for representation of the ontology is a meta-ontology of the form:

\[
O_M = \{ C, R, T, D, A, Ax \},
\]

where \( C = \{ C_1,..., C_n \} \) is a finite nonempty set of classes describing concepts of some subject or problem domain;

\[
R = \{ R_1,..., R_m \}, R_i \subseteq C \times C, R = \{ R_P \} \cup \{ R_A \} \cup R_A
\]

is a finite set of binary relations defined on classes (concepts):

\[
R_P \text{ is an antisymmetric, transitive, nonreflexive binary relation of inheritance defining a partial order on the set of concepts } C
\]

\[
R_A \text{ is a binary transitive relation of inclusion ("whole-part" relation),}
\]

\[
R_A \text{ is a finite set of associative relations,}
\]

\[
T \text{ is a set of standard data types,}
\]

\[
D = \{ d_1,..., d_k \} \text{ is a set of domains } d_i = \{ s_1,..., s_k \}, \text{ where } s_i \text{ is a value of the standard type string;}
\]

\[
TD = T \cup D \text{ is a generic data type, containing the set of standard data types and the set of domains;}
\]

\[
A = \{ a_1,..., a_n \}, \text{ where } A \subseteq C \times TD \text{ or } A \subseteq R_A \times TD \text{ is a finite set of attributes, i.e. binary relations like } a_i(C_j,td_k) \text{ or } a_i(R_j,td_k) \text{ describing the properties of concepts } C \text{ and relations } R_A;
\]
$F$ is a set of constraints imposed on values of attributes concepts and relations, i.e. predicates like $p_1(e_1, \ldots, e_m)$, where $e_k$ is either a name of an attribute ($e_k \in A$) or constant ($e_k \in t_j$, where $t_j \in TD$).

This formalism provides description of both concepts of its problem and subject domains and diverse semantic relations between them. It allows ordering concepts in a "generic-specific" hierarchy (with the help of the relation $R_T$) and supporting inheritance of properties through this hierarchy. A feature of the formalism is that under inheritance not only all attributes but all relations are passed from a parent class to a descendant class. The set of associative relations $R_A$ is defined by a user (developer of ontology). Availability of these relations allows us to organize a content-based search and navigation through the portal content. An important feature of relations $R_A$ is that they can have their own attributes which specify a link between its arguments.

Introducing formal descriptions of concepts of a subject domain in the form of classes of objects and relations between them, the portal ontology defines the structures for representation of real data and connections between them. According to this, the portal data themselves are represented as a set of linked heterogeneous information objects.

III. METHODOLOGY OF ONTOLOGY BUILDING

From the informal point of view, the portal ontology serves for representation of the concepts required for the description of both the problem domain of the portal and its area of knowledge. To make adjustment of knowledge portal and its maintenance easy, the portal ontology is built on the basis of the basic ontologies (Fig. 1) by means of its completion and evolution [4].

![Ontology of research activity](image)

**Fig. 1.** The basic ontologies of a knowledge portal.

Two ontologies were selected as basic, these are the ontology of research activity which constitutes the basis of the problem domain ontology of the portal and the ontology of scientific knowledge which is used as the basis for building a subject domain ontology (ontology of the area of knowledge of the portal).

The ontology of research activity plays a role of the problem domain ontology of the portal. Practically, it is a top-level ontology and includes the basic classes of concepts related to the organization of research activities, such as Researcher, Organization, Event, Activity, and Publication. This ontology also includes the class Information resource that serves for description of the information resources presented in the Internet.

The ontology of scientific knowledge states the main structures used for building the subject domain ontologies. In particular, this ontology contains the meta-concepts specifying the structures for the description of concepts of a specific field of knowledge, such as Subdivision of science, Research method, Object of research, Scientific result. Using these meta-concepts, we can describe divisions and subdivisions that are significant for a given science, determine classification of methods and objects of research, and describe the results of research activity.

The concepts of the basic ontologies are interconnected with each other by associative relations, selection of which is actualized taking into account not only preciseness and completeness of representation of the problem and subject domains of the portal, but also convenience of navigation through its content and information search.

The main associative relations are the following:

- "Scientific field" – links events, publications, organizations, researchers or information resources with subdivisions of science;
- "Describes" – links publication with a scientific result, method or object of research;
- "Uses" – links research method with activity, person or organization;
- "Investigates" – attaches some activity to the object of research;
- "Result of" – serves for linking scientific results with the activity;
- "Resource of" – links information resources with any concept of ontology.

The concepts of the ontology of an area of knowledge (the subject domain ontology) are realizations of meta-concepts of the ontology of scientific knowledge and can be ordered in the "generic-specific" hierarchy.

IV. TECHNOLOGY OF BUILDING KNOWLEDGE PORTALS BASED ON ONTOLOGIES

Based on the methodology considered above, the technology of building knowledge portals was developed. This technology is oriented to experts in a subject domain. It allows them to collect in the framework of a uniform information space an
extensive knowledge and data related to a certain subject domain.

The main components of the technology are a set of the basic ontologies, the methodology of ontology building, expert interface providing access to software tools supporting the ontology building and content management, as well as a user interface allowing one to perform search and navigation through the portal content.

The ontology editor serves for ontology management. It is implemented as a Web-application and is accessible to authorized users. The ontology editor was designed in such a way that it is easy to understand and use for experts in humanities which are not experienced in programming and computer science. In particular, to meet these requirements, we refused to use such popular means of ontology building as the Protégé editor [5].

Management of the portal content is performed with the help of data editor that allows one to create, modify and delete information objects (IO) and relations between them. Forms for input of such data are automatically generated on the basis of classes and relations of the portal ontology.

Description of information resources is an important component of the information content of the portal. According to the definition of the base ontologies, each resource corresponds to the concept of Information resource.

The set of attributes and relations of Information resource is based on Dublin Core standard [6] and includes the following units: “Title of the resource”, “Address in the Internet” (URL), “Subject of the resource”, “Resource type”, “Language” etc. The description of the resource includes an instance of this concept (an information object) and a set of instances of relations that link this IO with other IOs (persons, organizations, events, results of researches, etc.).

To make possible a distributed development of ontologies, the expert interface has a procedure for granting privileges to experts of different levels.

Content-based access to systemized knowledge and information resources included in the portal content is provided by advanced facilities for search and navigation. Their operation is based on ontology. Due to this fact, a scenario of user communication with the portal consists in selection of information objects of a required class of the ontology (with the help of either visualization facilities or search facilities), filtration of the list of such objects, detailed browsing of each IO and transition along the ontological links (relations) from this IO to other information objects, as well as browsing of information resources, the references to which are contained in this IO.

The suggested technology was used in the projects aimed at development of the scientific knowledge Internet portals on archeology and computational linguistics.

V. APPLICATION OF TECHNOLOGY OF BUILDING KNOWLEDGE PORTALS

A. Knowledge Portal on Archeology

A knowledge portal on archeology was developed in A.P. Ershov Institute of Informatics Systems in cooperation with the Institute of Archeology and Ethnography of the Siberian Branch of the Russian Academy of Sciences. It was built to solve the problem of systematization and integration of accumulated knowledge and information resources on archeology, as well as to provide the content-based access to them.

This portal is oriented to wide circle of users – from scientists and lecturers to students and schoolchildren interested in achievements of archeology.

The developed ontology on archeology includes four basic hierarchies:

Subdivisions of archeology hierarchy presents the structure of the main directions of scientific activity in archeology. Examples of subdivisions of archeology are General archeology, Field archeology, and Reconstructive archeology which, in turn, are subdivided into more particular subdivisions of archeology.

Objects of research hierarchy systematizes the objects of research in archeology. These objects have such properties as description of an object, date of discovery, accuracy of dating, etc. Examples of objects of archeology are Archeological culture, Historical Person, as well as Artifact, Complex, and Monument discovered during archeological excavations.

Research methods hierarchy serves for description of approaches, research methods, principles, technologies, and methods that are applied to a certain type of archeological objects.

Scientific results hierarchy is intended for description of the results of research activity in archeology, such as discoveries, hypotheses, new laws, theories, historical facts, etc.

It was found that meta-concepts presented in the ontology of scientific knowledge were not sufficient to present chronological and geographical features of archeological knowledge. Therefore, the ontology of scientific knowledge was complemented by new classes – Archeological period and Place which are typical for a historical science.

The class Archeological period serves for dating the objects of research and binding the subdivisions of science to time domains. Information objects corresponding to concrete archeological periods constitute a hierarchy of nesting and historical sequence are described by the time domain.

The class Place allows one to associate the subdivisions of science with geographical locations and to define historical and chronological connections between archeological periods and their geographical locations.

Thus, the portal content presents knowledge on the main subfields of archeology, its objects, methods and results of research, which are united in accordance with chronological
and geographical principles, supplied with text descriptions and references to publications. Information saved in the portal content allows one to organize the content-based access to resources which present data bases and electronic archives of graphic and bibliographic documents, reports about archeological activity, as well as descriptions of museum collections.

B. Knowledge Portal on Computational Linguistics

The knowledge portal on computational linguistics provides effective access to linguistic information resources. The users of this portal are researchers, lecturers, and students involved in this branch of science and specialists who develop systems for natural language processing, analysis and synthesis of speech.

The ontology of computational linguistics built on the basis of the ontology of scientific knowledge, plays a role of the ontology of portal’s area of knowledge. The ontology of computational linguistics includes about 200 concepts which are organized into five basic hierarchies: Subdivisions of science, Research methods, Objects of research, Subjects of research, and Scientific results.

Subdivisions of computational linguistics hierarchy is based on classification of the main theoretical directions of computational linguistics and determines significant divisions and subdivisions of computational linguistics. Divisions of computational linguistics are Machine translation, Text processing, Speech analysis and Speech synthesis, etc. These general divisions are divided into more specific subdivisions. For example, Machine translation includes Automatic and Automatized machine translation.

Objects of research hierarchy determines classification of the objects of research. A Discourse as the base object of research is considered as a form of natural language existence and use. In particular, phonetic, syntactic and other linguistic phenomena are taken into consideration, as well as such forms of discourse as Text and Speech.

Research methods hierarchy serves for systematized description of various models and methods that are applied in computational linguistics. Here the groups of methods, such as Linguistic methods, Computer Science methods, Mathematical methods, are distinguished.

Scientific results hierarchy serves for structuring and description of results of research activity. It includes such kinds of results as Technologies and Software products, Applied systems and Linguistic resources (Lexical Ontology, Dictionaries and Text Corpora).

Subjects of research hierarchy describes Basic processes and Tasks associated with functioning of units of language in the process of communication, as well as Applied processes and Tasks which are of practical value and satisfy certain social needs.

All hierarchies of computational linguistics are connected by means of associative relations. One part of these relations is inherited from the base ontologies, the other part of them is specific relations of a given subject domain.

As the portal is devoted to computational linguistics, its content presents, first of all, knowledge about the main division of computational linguistics, its objects and subjects of research, its models and methods. As for information resources, they describe the results of activity of organizations and researchers obtained within scientific and commercial projects. These results are technologies, software products, applied systems and traditional linguistic resources, such as vocabularies, linguistic databases, and text and speech corpora.

VI. CONCLUSIONS

The experience of building knowledge portals, providing the integration of knowledge and information resources on humanities and the content-based access to them, demonstrates the soundness and productivity of the proposed approach. The suggested methodology of ontology building, the main principle of which is to build the ontology of a specialized knowledge portal by means of completion and evolution of the basic ontologies, considerably simplifies creation and maintenance of such portals.

The formalism suggested for specification of ontologies is flexible enough for representation of concepts of humanitarian area of knowledge that have a complex structure, and diverse semantic relations between them.

The ontology editor developed on the basis of this formalism have proved to be a good representation and systematization knowledge facility which is intuitive and easy-to-use for experts in humanities. The ontologies developed in the frameworks of this approach are a good basis for systematization and integration of the knowledge and information resources on humanities, providing convenient navigation through them and content-based search in terms of the modeled area of knowledge.

The suggested approach to building knowledge portals based on ontology was successfully applied to the development of knowledge portals on archeology and computational linguistics.

REFERENCES