Semantic Library as a Tool of Defining a Scientific Subject Area

Olga M. Ataeva, Vladimir A. Serebryakov

Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, Moscow, 119333, Russia

Abstract

The paper considers an information system designed to represent the subject area associated with science and its features. Highlighted general concepts for the formal description of such a subject area in the knowledge base of the semantic library. The peculiarity of these areas is that the data structure is subject to frequent changes. Therefore, the tools of organizing knowledge, which is a semantic library, should be sufficiently universal and not require deep technical knowledge. The paper describes the functionality of the system and its use when setting up a subject area. For each area, the set of resources can differ both in format and in the set of the resources themselves. The set of concepts that form the description of the library's content should be so universal that it can be adapted to the needs of a particular area. Three levels of metadata are used to represent the data.

Keywords

Semantic library, ontology, knowledge representation

1. Introduction

Various researchers have dealt with the issues of the semantic organization of knowledge since ancient times. Libraries specialized in specific areas usually use their classifiers to organize their resources. This approach provides a more detailed analysis of the content of documents and the correlation of semantic concepts of the contents of the library with a certain direction of the specialized area of knowledge.

The accumulated data have become available to a wide range of users through the network, the functionality of digital libraries is becoming more and more diverse, satisfying *the information needs* of users.

The focus of the proposed work is subject areas related to science and their features. General concepts for their formal descriptions in the knowledge base are highlighted. The peculiarity of these areas is that the data structure is subject to frequent changes [1–4]. The main emphasis is placed on the presentation of a generalized model of a scientific subject area and its features, implementation in search engines and differences from classical approaches to information retrieval in scientific data sets.

New problems and challenges also relate to the representation of knowledge in the information environment for various fields of science using modern approaches. To ensure the consumption of scientific information at a new level, first of all, it is necessary to move to a semantically meaningful representation of scientific knowledge extracted from information in the digital environment.

To represent the data of the subject area, it uses metadata of three levels: (1) universal concepts without reference to the subject area, or metametadata; (2) concepts for describing a specific subject area or metadata, the definitions of which are given in terms of the first level; (3) application domain data as such, represented in terms of second level metadata. Based on this metadata, user interaction interfaces are configured for navigation, editing and information retrieval.

The main task of creating and describing a generalized representation of scientific knowledge for a certain area is to help experts in organizing knowledge and providing access to it [5–9]. At the same

SSI-2021: Scientific Services & Internet, September 20–23, 2021, Moscow (online)
EMAIL: oli@ultimeta.ru (O.M. Ataeva); serebr@ultimeta.ru (V.A. Serebryakov)
ORCID: 0000-0003-0367-5575 (O.M. Ataeva); 0000-0003-1423-621X (V.A. Serebryakov)

time, the means of organizing knowledge should be sufficiently universal and not require deep technical knowledge.

The task was to create such an information system that could take into account all the variety of different types of resources of a scientific subject area that can be stored in it, and at the same time support its terminological description. One of the main tasks to be solved in the context of the system is to provide the ability to integrate data from sources that support the semantic description of the data model. In fact, such a system should be a constructor with an adaptable stored data content model to create a digital library of any direction. An adaptable data model allows you to describe an arbitrary data model of the library content within a subject area, fixed in terms of a thesaurus.

A new generation information system should take into account the variety of types of resources in a scientific subject area and at the same time support its terminological description. The main tasks of such a system are to provide the possibility of integrating data from sources that support the semantic description of the data model, and the development of an ontological representation of the content of the subject area, which would allow describing any types of resources from the integrated sources. At the moment, the distribution kit of the semantic library has been implemented and ready to use. The following is a description of the main ideas of the data model and subsystems, which are presented in the distribution of the information system.

2. About the data model

In the information model of the semantic library, concepts were introduced to describe the contents of the library for a certain subject area [10–13]. These concepts allow you to construct a description of any type of information resources for this area. At the same time, according to the definition, information objects that are directly the contents of a library have a distributed nature, which means that data can come from various sources and aggregate information about an information object from various sources, directly saving data in the library itself or storing links to identical objects in sources data.

To describe the resources that make up the content of a specific subject area, concepts are used that are common to any of them. That is, the set of concepts that form the description of the library's content should be so universal that it can be adapted to the needs of a particular area.

The content of the library is closely related to the thesaurus, which maintains relationships of various types both between concepts and between concepts and information objects. This allows you to implement flexible custom search, the result of which will be a balanced list of objects in the subject area. Collections of a wide variety of resource types are defined based on the same thesaurus. This approach is extremely useful for creating separate custom collections.

In fact, the concepts are divided into three categories: the first includes definitions of the concepts of the content of the semantic library, the second category refers to the definition of the concepts necessary to support the terms in the domain thesaurus, and the third includes the definitions necessary to describe the processes of integration of the content of these resources [14–23]. Based on these definitions, the main processes are described, such as, for example, the integration of data from different sources, categorization / classification, mapping of different data source models to a given subject area, construction of equivalence classes, etc.

3. Architecture

Consider a formal description of the system that defines its goals, functions, externally visible properties and interfaces. It also includes a description of the components of the system and their relationships, along with the principles that govern its design, operation and possible subsequent development. This description includes software subsystems, visualized properties of those subsystems, relationships between subsystems, and restrictions on their use. Moreover, each subsystem can consist of several levels of abstraction, and each level can have its own architecture. Below is a list of the main subsystems:

- Subsystem for describing the content of the information system,
- Thesaurus control subsystem,

- Subsystem of automated data processing and presentation,
- Subsystem for the implementation of data integration tasks,
- Recommender subsystem.

Each of these subsystems is responsible for a specific functionality and uses its own subset of concepts from the information model.

4. Content description subsystem

Let's consider one of the subsystems that determines the basic settings of the system. The set of concepts that make up the information model of the Libmeta library content is responsible for the universality of defining the system content: *an information resource* and, which describe resource instances. *An information resource* is the main unit for describing the content of a library, and an *information object* represents instances of information resources. Each of them has its own unique identifier. In fact, the semantic meaning of *an information resource* is equivalent to the concept of an ontology class with some restrictions in its description. The structure of the description of information objects is determined by the concepts of a*ttribute* and a *set of attributes*, which are defined when describing the corresponding resource. An attribute is an element describing a property of a resource, and a set of attributes is defined as a collection of attributes of different kinds. The types of attributes are as follows: *attribute, file, object, numeric, text, string*. In addition to defining the range of values of an attribute, an important characteristic is its type and the definition of the number of its values. To describe a specific information resource, the concept of *an attribute value* is used, which is closely related to the concept of an *attribute* and is actually a container for storing specific values of an *information object* of a certain type.

These concepts provide a structured description of content and provide support for its adaptability. This approach also provides the description of specific resources and their objects in the form of RDF triplets and provides SPARQL an access point for publishing data in machine-readable formats.

In general, a specific implementation of the library content model can be based on some imported ontology, the classes of which are converted into resources, the properties are described in terms of LibMeta attributes, the attribute sets actually define the domains of the ontology properties. When building a library resource model based on this ontology, all URIs of properties, relations and classes of the selected ontology are saved. If necessary, when importing the selected ontology into the system, you can change the set of concepts by expanding or, on the contrary, reducing it by means of the system.

Of course, this way of mapping the ontology to the concepts of the LibMeta system does not preserve the entire possible list of restrictions imposed on the properties and classes of the ontology initially, but its structural part remains, which is sufficient for solving problems defined within the system.

Figure 1 shows the basic concepts used to construct a description of the subject area within this subsystem.

When describing *information resources* and determining a set of their attributes, *the types of attributes* that form the structural description of the resource play an important role. Attributes are divided into several overlapping types: *search, descriptive, administrative, identifying*. In the formation of search interfaces, it is *the search attributes* that play an important role, which are used when performing an attribute search by resource types. The result of such a search is objects, a short description of which is presented to the user through descriptive attributes.

In fact, within the framework of this subsystem, the initial configuration of the configuration of the library content and its interfaces for a specific subject area is performed. Figure 2 shows the sequence of user actions to configure the system.

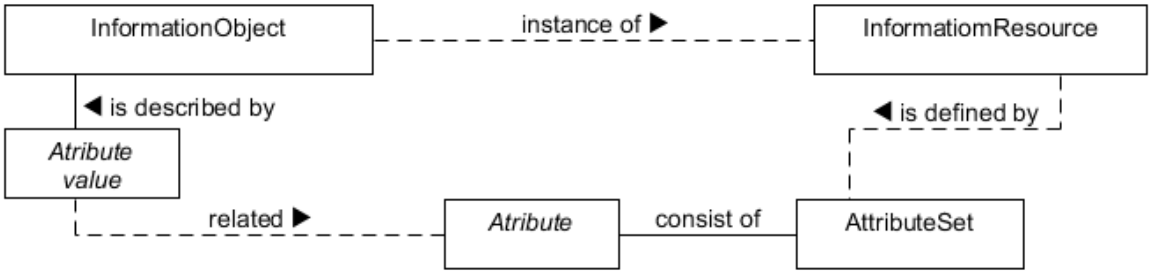

Figure 1: Basic concepts and their relations

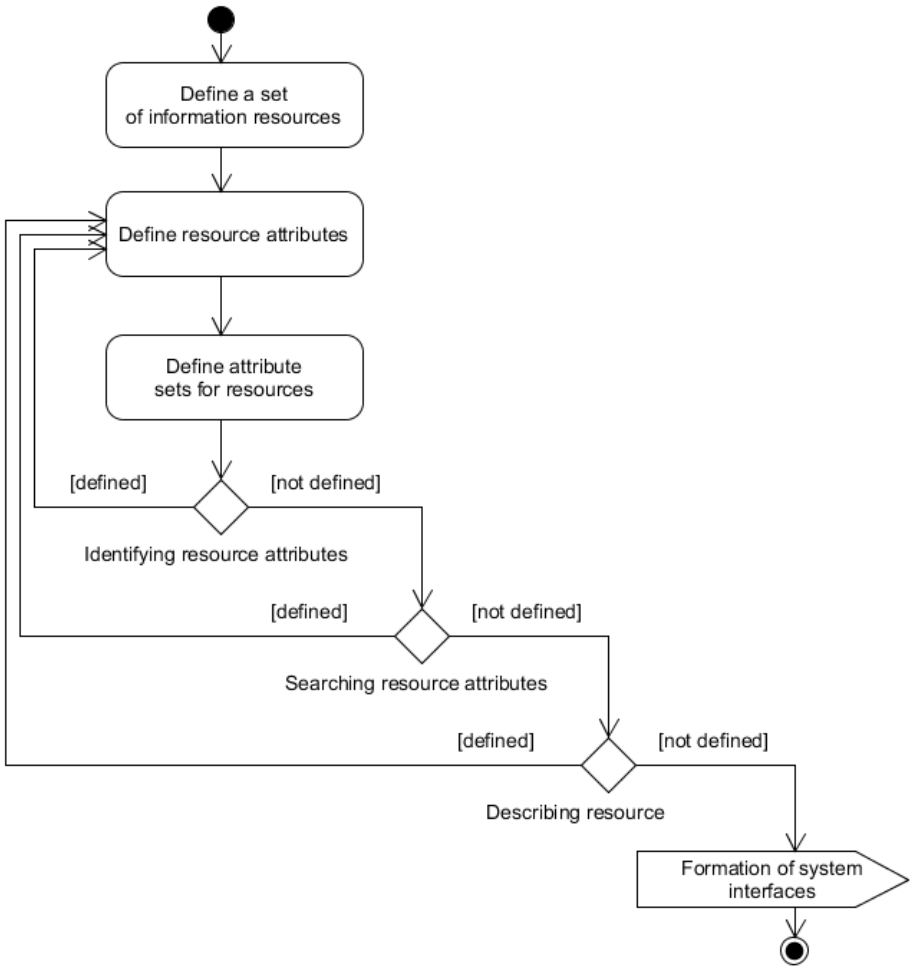

Figure 2: The sequence of user actions to configure the system

5. Basic functionality of LibMeta

Basic functionality of LibMeta:

- creation / viewing / editing of information resources and their structure;
- creation / viewing / editing of information objects and their structure;
- connecting data sources;
- loading data from connected data sources, which later become part of the library's content;
- creating / viewing / editing the structure of the thesaurus of the supported subject area;

- create / view / edit thesaurus concepts
- batch loading of data that make up the content of the library;
- attribute / semantic / full-text search and navigation through the available information objects of the system;
- attribute / semantic / full-text search by data sources;
- creating / viewing / editing collections of information objects;
- formation of a subject area ontology by describing the structure of information resources and thesaurus;
- provision of data constituting the content of the system in a machine-readable format;
- highlighting links between information objects and concepts of the thesaurus;
- support for semantic labels or folksonomy [24–26] to describe the thematic focus of information objects;
- creating / viewing / editing the user's area of interest;
- creation of a recommendation system:
  a) based on the description of the user's interests;
  b) based on the subject area thesaurus under consideration;
- support for user micro-thesaurus based on the domain thesaurus.

LibMeta functionality available to all public users:
- viewing information resources and their structure;
- viewing information objects and their structure;
- attribute / semantic / full-text search and navigation through the available system resources;
- attribute and semantic search over data sources;
- viewing public collections of information objects.

From the point of view of an authorized user, the semantic library additionally provides the following functionality:
- defining your micro-thesaurus as an extension of a certain node defined in the system of the main terminological thesaurus. It also provides support for the creation of so-called *annotation ontologies* or *user ontologies (folksonomy)*, which are a collective vocabulary of users, compiled as a result of the process of putting semantic labels on resources by them;
- defining your own collections of information objects;
- organization of joint thematic collections for user groups;
- attribute and semantic search on data sources with the ability to save search results;
- the user in the role of the system administrator has access to all the above-defined functionality and can use the additional functionality available only to him:
  a) can expand descriptions of resource types or create new ones at the request of users;
  b) can, at the request of users, include their resource objects in the public list of objects;
  c) for groups of users to make available the ability to edit certain types of resources or taxonomies;
  d) edit user groups and roles and the set of operations available to them;
  e) edit and configure the main terminological thesaurus and its links.

6. Conclusions

The description of the information system for the implementation of the functionality of the semantic library for a certain subject area is presented. Thus, subject matter experts get the opportunity to implement the main task of the library – the *semantic / intellectual* construction of the scientific knowledge space for a certain subject area. That is, endowing it with semantics by highlighting clearly intellectually meaningful connections, support for semantic markup. The main design tools are the ontology of the subject area, which allows you to meaningfully structure and ensure connectivity between the resources that are included in the scientific knowledge space of the subject area, and the use of unified terminological support in the form of a thesaurus of this subject area. To implement the

functions of openness of the scientific space of knowledge, the possibilities of integrating other data sources and the possibility of linking with their data were implemented. Providing functionality for collaborative work on the development of the scientific knowledge space increases the efficiency of research carried out in it and expands the possibilities for keeping it up to date.

References

[1]. Y. V. Leonova, A. M. Fedotov, Sozdanie prototipa sistemy upravleniya informacionnymi resursami, Vestnik Vostochno-Kazahstanskogo gos. Tekhn. Universiteta i zhurnala Vychislitel'nye tekhnologii, Kazahstan, (2018) 47–56.

[2]. M. V. Kulagin, A. S. Lopatenko, Nauchnye informacionnye sistemy i elektronnye biblioteki. Potrebnost' v integracii, Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii, 2001.

[3]. Y. I. Shokin, A. M. Fedotov, V. B. Barahnin, Problemy poiska informacii, 2010. URL: https://nsu.ru/xmlui/handle/nsu/161.

[4]. K. Börner, VIVO, A semantic approach to scholarly networking and discovery, volume 1 of Synthesis lectures on the Semantic Web: theory and technology, 2012.

[5]. N. B. Ngok, A. F. Tuzovskij, Obzor podhodov semanticheskogo poiska, Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki, 22, 2010.

[6]. Z. V. Apanovich, P. S. Vinokurov, T. A. Kislicina, Tools for Visual Analysis of Information Content of Portals Included in Linked Open Data Cloud, Conference “Digital libraries: Advanced Methods and Technologies, Digital Collections”, RCDL 2011, Voronezh, Russia, October 19–22, 2011, pp. 113–120.

[7]. E. A. Orobinskaya, A. Y. Doroshenko, Ispol'zovanie ontologij dlya avtomaticheskoj obrabotki tekstov na estestvennom yazyke, 2011. URL: http://repository.kpi.kharkov.ua/handle/KhPI-Press/14950.

[8]. B. V. Dobrov, N. V. Lukashevich, Tezaurus RuTez kak resurs dlya resheniya zadach informacionnogo poiska, Trudy Vserossijskoj Konferencii Znaniya-Ontologii-Teorii (ZONT-09), Novosibirsk, 2009. URL: http://ns.math.nsc.ru/conference/zont09/reports/93Dobrov-Lukashevich.pdf.

[9]. A. C. Ngonga Ngomo, et al, Sorry, i don't speak SPARQL: translating SPARQL queries into natural language, Proceedings of the 22nd international conference on World Wide Web, ACM, 2013, pp. 977–988.

[10]. V. A. Serebryakov, O. M. Ataeva, Osnovnye ponyatiya formal'noj modeli semanticheskih bibliotek i formalizaciya processov integracii v nej. Programmnye produkty i sistemy 4 (2015) 180–187.

[11]. O. M. Ataeva, V. A. Serebryakov, Personal'naya otkrytaya semanticheskaya cifrovaya biblioteka LibMeta, Konstruirovanie kontenta. Integraciya s istochnikami LOD. Inform. i eyo primen. 2, 11 (2017) 85–100.

[12]. O. M. Ataeva, Informacionnaya model' semanticheskoj biblioteki LibMeta. Programmnye produkty i sistemy 4 (2016) 36–44.

[13]. O. M. Ataeva, V. A. Serebryakov, Ontologiya cifrovoj semanticheskoj biblioteki LibMeta. Informatika i eyo primeneniya 12, 1 (2018) 2–10.

[14]. P. A. Lomov, M. G. Shishaev, Integraciya ontologij s ispol'zovaniem tezaurusa dlya osushchestvleniya semanticheskogo poiska. Informacionnye tekhnologii i vychislitel'nye sistemy 3 (2009) 49–59.

[15]. Y. Katsis, Y. Papakonstantinou, View-based data integration. Encyclopedia of Database Systems (2009) 3332–3339.

[16]. L. Xu, W. D. Embley, Combining the Best of Global-as-View and Local-as-View for Data Integration. ISTA 48 (2004) 123–136.

[17]. M. R. Kogalovskij, Metody integracii dannyh v informacionnyh sistemah. Institut problem rynka RAN, volume 74, 2010. URL: http://www.ipr-ras.ru/old_site/articles/kogalov10-05.pdf.

[18]. A. E. Karabach, Sistemy integracii informacii na osnove semanticheskih tekhnologij, Nauka, tekhnika i obrazovanie 2 (2014) 58–62.

[19].M. Lenzerini, Data integration: A theoretical perspective, Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, 2002, pp. 233–246.
[20].D. Calvanese, De G. Giacomo, M. Lenzerini, Ontology of Integration and Integration of Ontologies, Description Logics, 2001.
URL: http://www.diag.uniroma1.it/degiacom/papers/2001/CaDL01dl.pdf.
[21].N. F. Noy, Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record 33, 4 (2004) 65–70.
[22].L. Zhao, R. Ichise, Ontology integration for linked data. Journal on Data Semantics 4 (2014) 237–254.
[23].Le Hoaj, A. F. Tuzovskij, Razrabotka semanticheskih elektronnyh bibliotek na osnove ontologicheskih modelej, Trudy XV Vseros. nauch. konf. “Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii”, RCDL, 2013, pp. 143–151.
[24].A. Noruzi, Folksonomies:(un) controlled vocabulary. Knowledge Organization 33, 4 (2006) 199–203.
[25].L. Specia, E. Motta, Integrating Folksonomies with the Semantic Web. The Semantic Web: Research and Applications (2007) 624–639.
[26].T. Gruber, Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web and Information Systems 3, 1 (2007) 1–11.