ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2021, Vol. 42, No. 8, pp. 1920–1934. © Pleiades Publishing, Ltd., 2021.

Ontology Based Approach to Modeling of the Subject Domain "Mathematics" in the Digital Library

V. A. Serebryakov[1]* and O. M. Ataeva[1]**

(Submitted by A. M. Elizarov)

[1] *Dorodnitsyn Computing Centre of the Russian Academy of Sciences, Moscow, 119333 Russia*

Received March 31, 2021; revised April 19, 2021; accepted April 23, 2021

Abstract—The paper describes the approaches and methods to create a semantic library within the "Mathematics" subject domain. The theoretical background of the research involves the approach based on ontologies used when creating semantic libraries. The paper suggests a step-by-step description of the general ontology within the subject domain. The results obtained are well-grounded and contribute to distinct data integration.

DOI: 10.1134/S199508022108028X

Keywords and phrases: *semantic library, subject domain, ontology, thesaurus, data integration, data preprocessing, data cleaning, data model.*

1. INTRODUCTION

There is evidence to suggest that the approach towards information systems as applications focused on databases with rigid structures is rapidly changing under the weigh of information technology development. The information revolution led to new requirements for their functionality. On the one hand, there is a need for compliant data models when representing and interacting different systems. While on the other hand, there is a need for software flexibility in terms of supporting, applying and using data models that can change and adapt during their life cycle.

A modern information system should collaborate with various data sources. When describing the content of such sources and integrating with them, metadata plays the key role. Metadata based services supported by an information system could and should entail both data navigation mechanisms, a mechanism for querying and retrieving data, and the ability to annotate, profile, and personalize data.

Technologies that provide the related infrastructure for the development of modern information systems cannot be considered within the framework of only one scientific direction of information technology. Data modeling methods proposed within the Semantic Web paradigm are considered.

The problem of data semantics is to match the domain objects and the different sources that provide data for this subject domain. While the data sources were used separately in a closed environment, the problem of data semantics was not so relevant. The message is that when in a closed environment, the data content is interpreted by a limited number of users and applications based on a given data structure. The advent of the Internet as an open environment where both users and software agents constantly change the way the data is consumed became the crucial point in the development of both model-based technology and data integration. The Semantic Web paradigm that came later offers a variety of technologies to solve this problem, such as RDF, RDFS, and OWL. The key technology is ontologies (OWL). Their mission is to make data available for machine processing and remain its meaningful, allowing computers to perform data processing by programs of artificial intelligence.

There emerge new problems and challenges in knowledge representation within the information environment for various fields of science using modern approaches. There is a need to ensure the

*E-mail: serebr@ultimeta.ru

**E-mail: oli@ultimeta.ru

consumption of scientific information at a new level. Thus we first need to turn to a semantically significant way of scientific knowledge representation. The knowledge is extracted from information featured in a digital environment. It is clear that each scientific field has its own specifics. Current circumstances are characterized by multidisciplinary research and mutual scientific areas linkages. We need to develop universal approaches for the storage and presentation of scientific knowledge to achieve the ultimate goal. One of the research areas is reflected in the works [1–5].

2. PROBLEM STATEMENT

The information system of new generation is to both address various source types of the scientific subject domain and support its terminology. The main tasks of that system are to provide the possible integrating data from sources that support the semantic description of the data model, and to develop an ontological content representation of the subject domain, which would allow describing any types of resources from integrated sources. A prototype of such a system (hereinafter referred to as a "semantic library") was created by a group of authors [6–10]. Based on this approach, the experts created the semantic library "The Encyclopedia of Mathematics". The library combined several different sources and established links between them.

Thesauri and ontologies have been developed for the "Mathematics" subject domain [11, 12, 14, 21, 27, 28]. They accumulate data from different areas of mathematics. But despite the value of those decisions, a common approach to data modeling in this subject domain has yet to be developed.

A new common approach based on ontologies will make it faster and easier to integrate different data sources, allow for a more conscious search, link data from different sources, and enrich and complement existing information. The search in such an environment becomes more personalized, adapting to the user's profile. In this paper we consider an ontology-based approach to domain modeling [10].

3. SOURCES FOR MATHEMATICS

The data sources that were used in the work when forming the semantic library for the "Mathematics" subject domain are the "The Encyclopedia of Mathematics" [14], the ODE thesaurus [12, 28], the special function dictionary, the equations of mixed type dictionary [19, 27] industry classifiers [15], mathematical articles in the tex, Dbpedia [16], Mathnet [17] formats, the English version of "The Encyclopedia of Mathematics" [18]. We will describe some features of these data sources.

1. The Encyclopedia of Mathematics is a five-volume Soviet encyclopedic publication devoted to mathematical topics. It is a fundamental illustrated publication on all major branches of mathematics, consisting of more than 6 thousand articles. The book was published in 1977–1985. Later, the Encyclopedia of Mathematics was digitized. The digitized articles are unstructured texts and formulae in the form of pictures that do not contain any links to related articles of the encyclopedia or other sources, and do not have references to the mathematics section. These disadvantages make it insufficiently suited for the Internet users to have it within the framework of an digital library [14].

2. The ODE Thesaurus—is a type of thesaurus which contains a lexical and semantic index, a list of terms, and literature. What counts most is that it contains both the concepts and terms themselves, and links to publications that introduce/define these concepts, their mathematical notation [12, 28].

3. The special function dictionary and the equations of mixed type dictionary both were compiled by experts in mathematical physics. It is a collection of basic formulas with guidelines and explanations [19, 27].

4. Industry classifiers (MSC and UDC) are hierarchical structures with horizontal binding, they are recognized in the professional community and provide a more detailed analysis of the document content. They also relate semantic concepts of content with a certain direction of a knowledge field [15].

5. Mathematical articles are texts in *tex* format from various journals

6. Dbpedia provides access to structured information from Wikipedia. One of the most well-known examples of fulfilled concept with linked data within the Semantic Web [16].

7. MathNet is a Russian mathematical portal that provides various opportunities to search for information about mathematical life in Russia [17].

8. English version of "The Encyclopedia of Mathematics" — In 1987, The Encyclopedia of Mathematics was translated into English and supplemented with about two thousand new articles. To date, the electronic version of the English Encyclopedia of Mathematics is supported by the international publishing house Springer (Luxembourg) and is available online [18]. The Encyclopedia of Mathematics features the articles with formulas in the TEX format, suitable for machine processing, it also provides links to related articles in the encyclopedia. Each article is associated with the MSC index [15], which is used to classify sections of mathematics. Together, this metadata opens up a wide range of opportunities for users to search for articles based on their interest and study related topics.

4. DATA PREPROCESSING AND PREPARATION FOR LOADING

One of the unavoidable steps in preparing data for loading its text formats into an already prepared data infrastructure is preprocessing and cleaning up this data. In this case, the data was provided in files in tex format for two sources: *articles in tex format*, and *the ODE thesaurus*.

Since the files were designed in different styles and the commands were given different names, it was necessary to replace all the author's tags with standard ones and to clear the documents of special characters and unknown tags. Yet we failed to completely avoid manual data processing, but at least it was possible to minimize it.

4.1. Article Preprocessing

The preprocessing block is written in the Python programming language together with the integration of the open-source library TexSoup (version 2015). It is divided into the following blocks: clearance of a document, converting the article into a tree representation, processing all the tree nodes, writing the corrected document. Figure 1 shows the main stages of text processing.

4.2. ODE Preprocessing

The thesaurus files contained both valuable information, many service and auxiliary symbols and whole words. The text analysis via regular expressions was the most effective method to extract informative data. Figure 2 below shows a fragment of a thesaurus file containing information about indexes and concept names.

Pairs of concepts connected by associative or generic relations were extracted separately. The links were extracted in several stages. First, we extracted all the informative strings, then came the related concepts and types of links, then we separately processed the information about all the concepts that are related to the main concept.

A similar process with regular expressions was used to extract formulas and notes from the corresponding sections of the documents. The data was then described with ontology terms and prepared for uploading. The final result is available on the library's website. A fragment of the thesaurus prepared for upload and described with general ontology terms of semantic library is shown in Figure 3. The structure of this ontology will be discussed below.

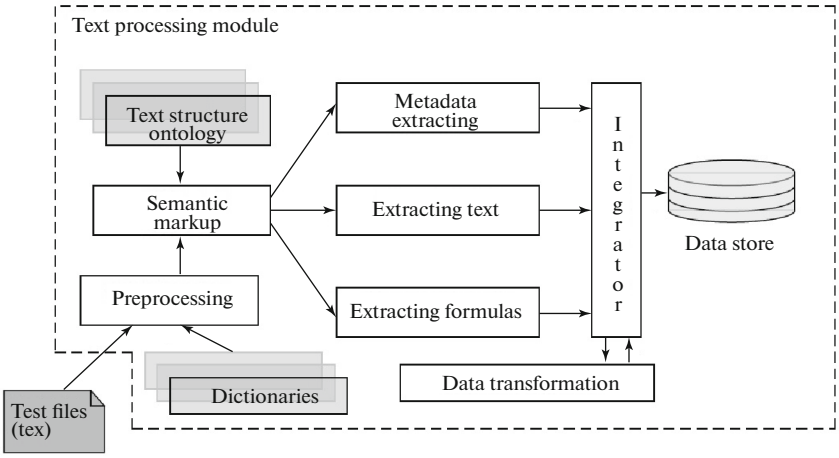


Fig. 1. The main stages of word processing.

```
\sk9pt\noi{\bf |DE0001|}$\hpa$ОДУ первого порядка

\sk3pt\noi{\bf |DD0002|}$\hpa$ОДУ первого порядка в полных
дифференциалах и метод
интегрирующего мно-\lb $\hphantom{\mbox{\bf
|DE0002|}}\hpa$жителя ОДУ первого порядка

\sk3pt\noi{\bf |NOD0097|}$\hpa$Локальное свойство существования
и единственности
решения ОДУ первого\lb $\hphantom{\mbox{\bf
|NOD0002|}}\hpa$порядка

\sk3pt\noi{\bf |RDE0162|}$\hpa$Фазовый портрет в окрестности
<<точки покоя системы двух линейных одно-\lb
$\hphantom{\mbox{\bf
|RDE0002|}}\hpa$\rzs{родных} \rzs{уравнений} с \rzs{постоянными}
\rzs{коэффициентами} (\rzs{характеристические}\lb
$\hphantom{\mbox{\bf |RDE0002|}}\hpa$корни: $\lm_1=\lm_2=0;$
матрица нулевая>>

\sk3pt\noi{\bf |DM0111|}$\hpa$Н.\,П.~Еругина метод нахождения
особого решения ОДУ первого порядка
\rzs{RDE0130}, \rzs{RDE0129}, \rzs{RDE0156}, \rzs{RDE0126},
\rzs{RDE0124}, \rzs{RDE0118}, RDE0122, RDE0121

DE0137,\hfill RDE0143,\hfill RDE0140,\hfill DE0142,\hfill
RDE0110,\hfill DE0174,\hfill DE0136

\rzs{RDE0132}, \rzs{RDE0130}, \rzs{RDE0156}, \rzs{RDE0158},
\rzs{RDE0126}, \rzs{RDE0124},\lb RDE0118, RDE0122, RDE0121
```

Fig. 2. Fragment of the thesaurus source file.

4.3. Preprocessing the Encyclopedia of Mathematics

At the preprocessing stage we included information about the articles belonging to the mathematics section, placed cross-references between the articles and determined computer-readable formulas related to the articles. This helps both build queries to the Encyclopedia of Mathematics, and further integrate with other knowledge bases.

We used the data contained in the English version of the encyclopedia (The Encyclopedia of Mathematics) to achieve this goal. Particularly, pointers to MSC sections and formulas in TEX format from articles. We need to compare the Russian and English version of articles, featured in the Encyclopedia of Mathematics to use them. The cross-references between the articles were carried out via the methods of semantic annotation [22–25]. Thus, the task of pre-processing was to perform the following steps:

1. Comparing Russian and English version of articles, featured in the Encyclopedia of Mathematics.

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:lbm="http://libmeta.ru/" xmlns:core="http:
 ▼<lbm:Concept rdf:about="http://libmeta.ru/thesaurus/concept/DE0008">
    <lbm:thesaurus rdf:resource="http://libmeta.ru/thesaurus/ODU"/>
    <lbm:code>DE0008</lbm:code>
    <core:url>http://libmeta.ru/concept/show/437</core:url>
    <lbm:descriptor>Абеля ОДУ второго рода</lbm:descriptor>
    <lbm:comment/>
    <lbm:conceptGroup rdf:resource="http://libmeta.ru/thesaurus/group/main"/>
   ▼<lbm:properties>
    ▼<lbm:property rdf:resource="http://libmeta.ru/thesaurus/attribute/includeFormula">
       <lbm:value rdf:resource="http://libmeta.ru/thesaurus/concept/DE0008_1"/>
     </lbm:property>
    ▼<lbm:property rdf:resource="http://libmeta.ru/thesaurus/attribute/maths">
     ▼<lbm:value>
        <![CDATA[ <math xmlns="http://www.w3.org/1998/Math/MathML" display="block"><mfenced close="]" open="[">
        <mspace width="0.167em"/><mi>y</mi><mo>+</mo><msub><mi>g</mi><mn>0</mn></msub><mfenced close=")" open="(
        </msup><mo>=</mo><msub><mi>f</mi><mn>2</mn></msub><mfenced close=")" open="("><mi>x</mi></mfenced><mspac
        </msub><mfenced close=")" open="("><mi>x</mi></mfenced><mspace width="0.167em"/><mi>y</mi><mo>+</mo><msu
        </math> (1) где <math xmlns="http://www.w3.org/1998/Math/MathML"><msub><mi>g</mi><mn>1</mn></msub><mo>,<
        width="0.167em"/><msub><mi>f</mi><mi>i</mi></msub></math> <math xmlns="http://www.w3.org/1998/Math/MathM
        <mn>0</mn></mrow><mrow><mspace width="0.167em"/><mn>1</mn></mrow><mrow><mspace width="0.167em"/><mn>2</m
       </lbm:value>
     </lbm:property>
    ▼<lbm:property rdf:resource="http://libmeta.ru/thesaurus/attribute/literature">
       <lbm:value rdf:resource="http://libmeta.ru/resource/publications#odu#25012"/>
     </lbm:property>
    ▼<lbm:property rdf:resource="http://libmeta.ru/thesaurus/attribute/note">
     ▶<lbm:value>
       ...
       </lbm:value>
     </lbm:property>
   </lbm:properties>
  ▼<lbm:familyRelation type="http://libmeta.ru/relation/family#association">
     <lbm:value rdf:resource="http://libmeta.ru/thesaurus/concept/DE0010"/>
   </lbm:familyRelation>
  ▼<lbm:familyRelation type="http://libmeta.ru/relation/family#association">
     <lbm:value rdf:resource="http://libmeta.ru/thesaurus/concept/DE0009"/>
   </lbm:familyRelation>
 </lbm:Concept>
</rdf:RDF>
```


Fig. 3. Fragment of the thesaurus prepared for loading in terms of the general ontology of the semantic library.

2. Providing articles featured in the Encyclopedia of Mathematics with annotations-links to other articles of the encyclopedia.

3. Assigning the articles with the MSC indexes which is similar to the Encyclopedia of Mathematics.

4. Comparing the articles, featured in the encyclopedia of formulas in TEX format obtained from the corresponding articles from the Encyclopedia of Mathematics.

5. Comparing lists with the content similar to articles.

6. The ability to upload data into a semantic library.

The structure of the concepts, featured in the Encyclopedia of Mathematics does not have a hierarchy as such, yet due to the concept related MSC codes, we were able to distinguish thematically related terms used in separate sections of mathematics. We identified the persons mentioned in the articles and put down the connections between the concepts and persons. We separately indexed the formulas. We also associated a set of corresponding formulas with each concept, if possible.

5. THE ONTOLOGICAL APPROACH

The set of sources may differ both in format and in the sets themselves for each subject domain. The set of concepts that form the description of the library content should be so universal that it can be adapted to the needs of a particular field. One of the main tasks solved within the library is the data integration from various sources. That is why these integration tools can be adapted to the conditions of any subject domain without reference to its specifics. This approach is the basis for the semantic library LibMeta [6–10].

The concepts that make up the library's ontology are conditionally divided into those intended for:

– domain content descriptions,

– creating a thesaurus for any subject domain,
– thematic collections descriptions,
– describing the task of integrating semantic sources.

Semantically significant connections are defined between these groups of concepts. Let us consider the basic definitions necessary to describe an ontology.

Definition 1. The content of a library $C = \langle IR, A, IO\rangle$ *is defined by the types of its data sources, described by the associated sets of attributes* A *and a set of inputs defining* IO *information objects, which are directly objects stored in the library.*

Definition 2. The library thesaurus $TH = \langle T, R\rangle$ *is defined by the terms* T*, the relationships* R *between them. The set of terms* T *that make up the domain description is in unvarying sequence.*

Definition 3. Semantic labels $M = \{m_i\}$ *of an information object are terms that are not included in the thesaurus but are necessary for specifying the subject of the information object. Semantic labels are not related, (unlike thesaurus terms), to each other or to some thesaurus terms, but allow for an additional subject division of information objects within the subject domain.*

Definition 4. The task of library data integrating $IT = \langle DS, R, A, M, D, D_S\rangle$ *with external* DS *sources is defined by the types of library resources and their set of attributes* A*, the mapping of* M *resources* R *to the data source schema* S*, and the data set of the source library associated by this mapping with the data* D_S *of the source.*

Definition 5. *A collection of information objects* $C = \langle IO, T, M, DS\rangle$ *is a set of objects combined on the basis of an entirety of features:*

1. *by their thesaurus term from the subject domain,*
2. *by semantic labels,*
3. *by the data source that the objects came from.*

The collection can include objects of various resource types specified in the content description.

Definition 6. Semantically significant library relationships $P = P_i$ *are the relationships defined between the library content, its subject domain (thesaurus), semantic labels, and data source objects. The authors highlight the following basic linkage:*

- $P_1(t, io)$*—thesaurus term-information object,*
- $P_2(io, t)$*—information object—thesaurus term,*
- $P_3(r, s)$*—data source—a source objects class, where a data source is a general definition for information objects stored in the system. Thus, information objects are instances of data sources,*
- $P_4(a, s_a)$*—data source attribute-source class property,*
- $P_5(io, o_s)$*—information object—a class instance from a data source,*
- $P_6(m, io)$*—semantic label—information object,*
- $P_7(io, m)$*—information object—semantic label.*

Based on the introduced explicit relations, we can determine the relations, which we will call **implicit meaningful relations** *(that is, set according to some pre-defined rules) between semantic labels and thesaurus terms and both the library objects and instances of related data from sources:*

- $P_8(m, t) \leftarrow P_6(m, io) \wedge P_2(io, t)$ *semantic label—information object—thesaurus term,*
- $P_9(t, m) \leftarrow P_1(t, io) \wedge P_7(io, m)$ *thesaurus term—information object—semantic label*,
- $P_{10}(m, o_s) \leftarrow P_6(m, io) \wedge P_5(io, o_s)$ *semantic label—information object—class instance from a data source,*
- $P_{11}(t, o_s) \leftarrow P_1(t, io) \wedge P_5(io, o_s)$ *thesaurus term—information object—class instance from a data source.*

To represent an ontology in OWL, classes, class properties, and individuals are used. In OWL terms, P_1 is *inverse* ofP_2, P_6 is *inverse* of P_7, P_8 is *inverse* of P_9, P_{10} is *inverse* of P_{11}. In this case, the rules for implicit relations are set using SWRL rules. The SWRL language, as an extension of OWL, helps to describe the abstract mechanism of operating with subject domain objects and regularities. The rules allow to deduce new facts from existing statements, increasing the efficiency of the subject domain description.

In compliance with the definitions, the ontology classes necessary for domain modeling were introduced:

1. ***IResource*** (library information resource), which contains general information about the resource type, name, URI, and information about the attribute set used to describe the structure of the resource.

2. ***IObject*** (library information object), which is an instance of a resource with the composition of the attributes corresponding to a set of attributes of the associated resource. To describe the corresponding values for an information object, there is a multivalued value property, values of which are instances of the ***AttributeValue*** helper class that contains information about the specific value of the object as well as the corresponding attribute.

3. ***Attribute*** is a superclass for classes of elements to describe composite objects of the subject domain:

   ResourceAttribute is a class to describe elements of the subject domain resource structure.

   ThesaurusAttribute is a class to extend the structure of the thesaurus elements description.

4. ***AttributeSet*** is a set of attributes that groups attributes that correspond to a single resource.

5. ***Taxonomy*** is a superclass to describe linear dictionaries, and classifiers, represented by the ***Vocabulary*** and ***Classifier*** classes correspondingly.

6. ***Thesaurus*** contains general information about the thesaurus: title and authors, and other information about thesaurus structure. The presence of this entity allows you to upload finished thesauruses without mixing them with those that may already be in the system.

7. ***Concept*** is an entity containing information about the concepts of the thesaurus.

8. ***Relations*** is a superclass for the relations classes that define the structure of the dictionary: ***HierarchicalRel*** are the hierarchical relations, ***FamilyRel*** are the horizontal relations.

9. ***PrefferedTerm*** are the descriptors of the concept. Each concept corresponds to a single descriptor in each language.

10. ***NonPrefferedTerm*** this includes synonyms.

11. ***SemanticTag*** is a class of semantic labels.

12. ***DataSource*** is a data source with a semantic wrapper (for example, a data source from LOD)

13. ***ResourceMapping*** is a class that contains information about the information resources of the library displayed for the data source.

Figure 4 below shows a part of the class diagram of the first-level ontology, Figure 5shows a description fragment of the same ontology properties in RDF/XML format.

6. THREE-LEVEL ONTOLOGY

The LibMeta ontology [6–10] uses three levels of metadata to represent subject domain data:

1. universal concepts without reference to the subject domain or metadata;

2. concepts for describing a specific subject domain or metadata, which definitions are set in the first-level terms (metametadata);

3. subject domain data as such, represented in terms of second-level metadata.

In such an ontology, concepts that are related to high-level ontologies and are not related to the specifics of a particular subject domain are used at the top level. At the second level, we describe the concepts of a specific subject domain as instances of first-level classes, i.e., for example, a specific thesaurus, specific types of information resources, types of data sources, etc.

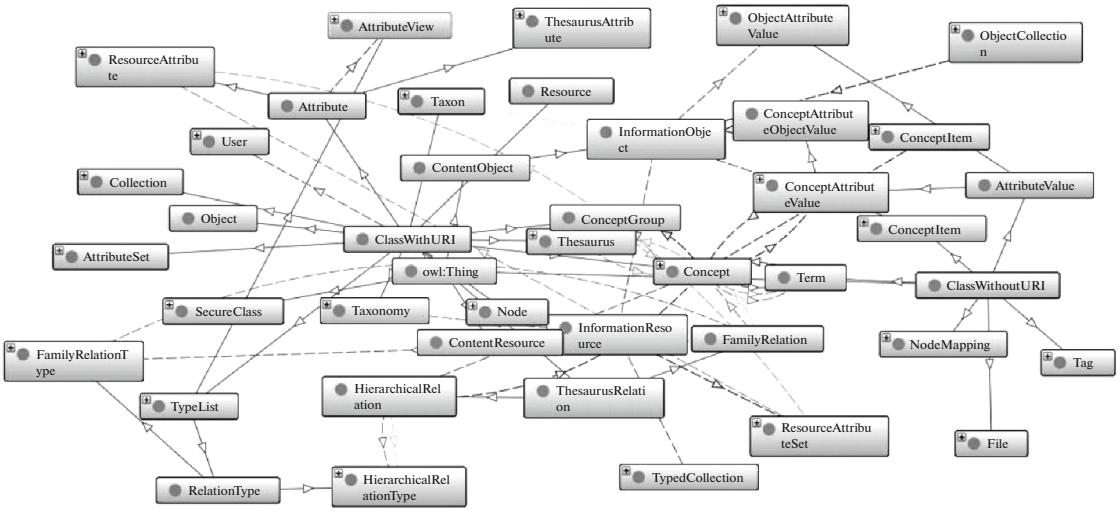


Fig. 4. Part of the first-level ontology class diagram.

Second-level concepts are used as class definitions at the third level when uploading data to an ontology with instances of second-level classes data.

At the same time, if at the second level the newly introduced concepts are instances of the first level designated resources, then when uploading data to the ontology, we use them as classes to describe the data. Considering instances as classes is called "metamodeling." And although even the direct semantics of the OWL2 ontology language, which is used to describe ontologies, do not allow such metamodeling. This means that when an instance identifier occurs in a class axiom, it is treated as a class, and when the same identifier occurs in a separate statement, it is treated as an instance.

So, when constructing an ontology of a specific subject domain, we, in fact, construct a three-level ontology, in which instances of the first level are high-level concepts, with the second level containing concepts of a specific subject domain. When uploading data to the ontology we use the first level terms to define the third level classes.

6.1. Ontology Building for the Subject Domain "Mathematics"

The data sources discussed above can be divided by type into three groups: texts (mathematical articles) that directly represent data for the subject domain, taxonomies (The Encyclopedia of Mathematics, ODE Thesaurus, Special Function Dictionary, Industry classifiers (MSC and UDC) used for terminological support of the subject domain, external sources (Dbpedia, Mathnet, English version of The Encyclopedia of Mathematics). External sources provide additional information about the subject domain data by establishing relations between the data in the subject domain modeled part and the sources content.

Each group represents the following types of resources: Data about a *person* and a *journal*, is extracted from mathematical articles that represent an obvious type of resource such as a *publication*. Also, *formulas* are extracted from the articles, and some structural elements, such as *theorems, lemmas*, etc. are highlighted.

Taxonomies distinguish dictionaries, classifiers, and thesauruses. The dictionary is a linear structure, the classifier is a hierarchical structure which can be used to support horizontal relations. The structure of an element in dictionaries and classifiers differs slightly, and usually such attributes as *code, name, language*, and *note* are enough to describe it. Thesauri are vertically and horizontally associated sets of concepts. Each relation has its own type (genetic relation, association, etc.). The structure of concepts from thesaurus to thesaurus can change significantly, yet there are common attributes such as *descriptor, non-descriptor*, and *synonym*.

External sources are a separate type of resources named *data source*. Each has a semantic layer which represents its data model. Each data source can be associated with any of the resource types described above.

Since the list of resource types and their structure may change depending on the incoming data, you need to provide additional structures to configure their descriptions. Some description details are omitted here and further not to encumber the article, yet this does not affect its correctness and comprehension.

Level 1. Thus, we have identified the following general types of resources to describe our subject domain

• Information resources (***IResource***)
• Taxonomy (***Taxonomy***)
– Classifier (***Classifier***)
• Thesaurus (***Thesaurus***)
• The concept of a thesaurus (***Concept***)
• Relations of the thesaurus (***Relations***)
v External sources (***DataSource***)
• Mapping (*textbfResourceMapping*)
• Attribute (***Attribute***)
– Thesaurus attribute (***ThesaurusAttribute***)
– Information resource attribute (***ResourceAttribute***)
• Multiple attributes (***AttributeSet***)

These resource types correspond to the first level of the LibMeta ontology (the metametadata level) and to the ontology classes specified in parentheses.

Level 2. Based on the first level concepts, we introduce concepts to describe our subject domain (metadata level) building on the listed data sources.

• Information resources (***IResource***)
– *Person*
– *Publication*
– *Formula*
• Classifier (***Classifier***)
– *MSC*
– *UDC*
• Thesaurus (***Thesaurus***)
– *The Encyclopedia of Mathematics*
– *ODE Thesaurus*
• External sources (***DataSource***)
– *English version of The Encyclopedia of Mathematics*
– *MathNet*
• Mapping (***ResourceMapping***)
• *English version of The Encyclopedia of Mathematics—The Encyclopedia of Mathematics*
– *MathNet-Person*
– *MathNet-Publication*
• Thesaurus attribute (***ThesaurusAttribute***)
– *Mathematical notation*
– *Formula*
– *MSC*
– *UDC*
– *See also*
• Information resource attribute (***ResourceAttribute***)
– *Annotation*
– *Formula*
– *Full name*
– *Title*
• Multiple attributes (**AttributeSet**)
– *Multiple attributes of the ODE thesaurus concept*

Fig. 5. Fragment of description of ontology properties in rdf/xml format.

– *Multiple attributes of The Encyclopedia of Mathematics*
– *Multiple person attributes*
– *Multiple formula attributes*
– *Multiple publication attributes*

The concepts of the modeled subject domain, which correspond to the second level of the Libmeta ontology, are italicised.

For example, for the ability to upload data into a semantic library the description of the Encyclopedia of Mathematics with a three-level ontology terms includes such concepts as *Thesaurus, Concept, Term, HierarchicalRelation, FamilyRelation* [9, 13]. Also, the thesaurus description to upload the Encyclopedia of Mathematics is expanded additionally with the help of such attributes as: *formula, person, UDC code, MSC code, reference link* (to the English version of the concept).

Level 3. If the second level is a modeling of the subject domain data structure within the terms of the first level, then the last level features the data in the described format published. You can see the result of publishing subject domain data on the project's website.

Figures 6 and 7 show examples of a specific information resource and information object description within the terms of this ontology according to **Definition 1**.

7. THE REQUIREMENTS TO SEMANTIC DIGITAL LIBRARY

The content of the semantic library should feature versatility, structure, adaptability to be supported by a three-level ontology and modeling tools in the LibMeta system. The versatility provides a description of its resources and objects types, regardless of the subject domain and the users interest area. The structure of the description provides links between different types of resources both inside and outside the system. The adaptability of the resource description allows for adding new properties and links in the process of system development and customizing of user interface to reflect perspective changes [20, 21].

In fact, LibMeta provides the feature set of constructing the space of subject domain scientific knowledge within the library. At the initial stage of installing the system, it only requires configuring the system for a specific subject domain.

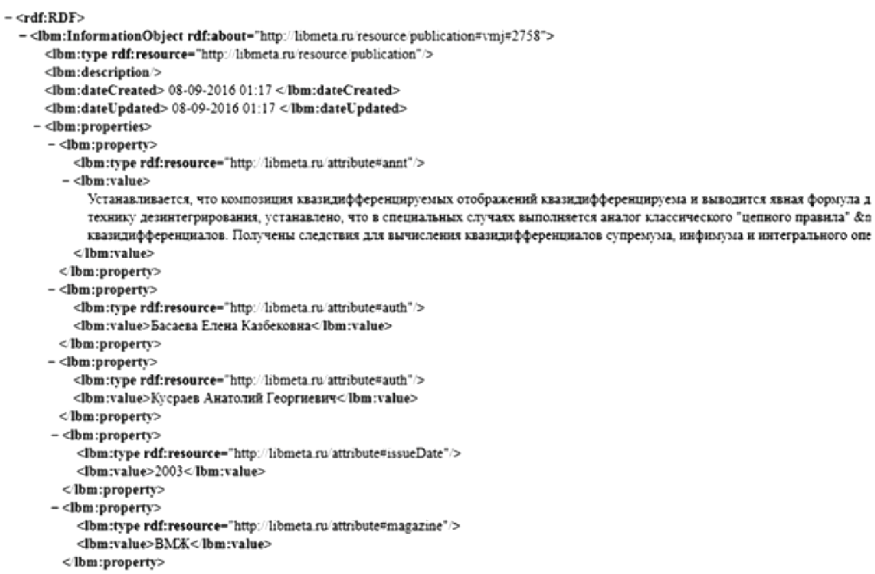
Here are the main types of tasks that are implemented in a semantic library that allows you to design a subject domain based on a three-level ontology:

- the information system content description;
- implementation of data integration tasks from external sources;
- collection support;
- search and navigation through system objects;
- user support.

The Figure 8 shows a set of subsystems that implement an information system feature set, depending on the level of ontology concepts used. At each level, the user's level of competence determines the access to feature set, as illustrated in Fig. 8.

```
- <rdf:RDF>
  - <lbm:InformationResource rdf:about="http://libmeta.ru/resource/person">
      <lbm:title>Person</lbm:title>
      <lbm:label>Персона</lbm:label>
      <lbm:description>Ресурс соответствующий персонам</lbm:description>
    - <lbm:properties>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#activity"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#additionEmployer"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#additionPosition"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#address"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#bio"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#dateOfBirth"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#dateOfDeath"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#employer"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#first"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#keywords"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#last"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#middle"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#placeOfBirth"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#position"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#seeAlso"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#source"/>
        <lbm:property rdf:resource="http://libmeta.ru/attribute#email"/>
      </lbm:properties>
      <lbm:dateCreated> 06-09-2016 13:32 </lbm:dateCreated>
      <lbm:dateUpdated> 06-09-2016 13:32 </lbm:dateUpdated>
    </lbm:InformationResource>
  - <lbm:InformationResource rdf:about="http://libmeta.ru/resource/publication">
      <lbm:title>Publication</lbm:title>
```


Fig. 6. An example of the description of information in terms of ontology.



```
- <rdf:RDF>
  - <lbm:InformationObject rdf:about="http://libmeta.ru/resource/publication#vmj#2758">
      <lbm:type rdf:resource="http://libmeta.ru/resource/publication"/>
      <lbm:description/>
      <lbm:dateCreated> 08-09-2016 01:17 </lbm:dateCreated>
      <lbm:dateUpdated> 08-09-2016 01:17 </lbm:dateUpdated>
    - <lbm:properties>
      - <lbm:property>
          <lbm:type rdf:resource="http://libmeta.ru/attribute#annt"/>
        - <lbm:value>
            Устанавливается, что композиция квазидифференцируемых отображений квазидифференцируема и выводится явная формула д
            технику дезинтегрирования, устанавлено, что в специальных случаях выполняется аналог классического "цепного правила" &n
            квазидифференциалов. Получены следствия для вычисления квазидифференциалов супремума, инфимума и интегрального опе
          </lbm:value>
        </lbm:property>
      - <lbm:property>
          <lbm:type rdf:resource="http://libmeta.ru/attribute#auth"/>
          <lbm:value>Басаева Елена Казбековна</lbm:value>
        </lbm:property>
      - <lbm:property>
          <lbm:type rdf:resource="http://libmeta.ru/attribute#auth"/>
          <lbm:value>Кусраев Анатолий Георгиевич</lbm:value>
        </lbm:property>
      - <lbm:property>
          <lbm:type rdf:resource="http://libmeta.ru/attribute#issueDate"/>
          <lbm:value>2003</lbm:value>
        </lbm:property>
      - <lbm:property>
          <lbm:type rdf:resource="http://libmeta.ru/attribute#magazine"/>
          <lbm:value>ВМК</lbm:value>
        </lbm:property>
```


Fig. 7. An example of a description of an information object in terms of ontology.

8. DATA SOURCES INTEGRATION

The semantic data source, by our definition, represents not only the data itself, but also contains a semantic layer in which the data model is described. Such sources, for example, include all sources of Linked Open Data, the core of which is the Dbpedia project.

Information resources of the system are aligned to the data sources, meanwhile the ratio of the resource attribute set and the properties of the resource from the data source is established. This provides ground for us to generate SPARQL queries to data sources to extract specific information. In this case, the user operates with the typical kind of search options, avoiding the need to write the queries themselves.

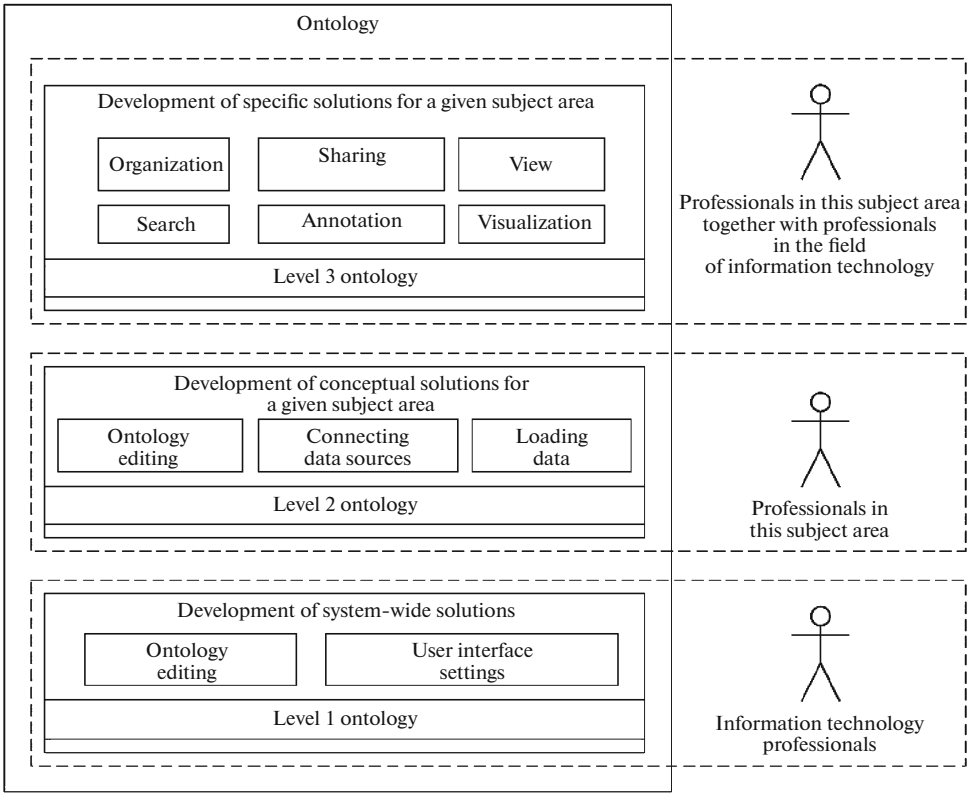


Fig. 8. A set of subsystems that implement the functionality of the information system, depending on the level of ontology concepts used.

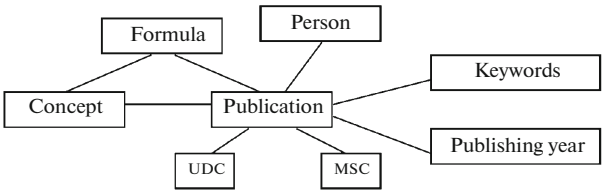


Fig. 9. Formula linkage network.

Using the MathNet example, we will describe a data integration from a semantic source. A database which data model represented person information within the terms of FOAF ontology provides person data from MathNet. At the second level of the ontology, the classes mapping and the corresponding resources attributes in the source was defined within the terms of FOAF model and the semantic library model. According to them, at the third level of the ontology, connections were formed at the level of instances that make up the semantic library content. In particular, there were added links to the MathNet person pages using the *see also* attribute.

9. INFERENCE RULES

The mathematical tool behind the descriptive logics, on which ontologies are based, provides the means to logically infer new facts from existing ones. Logical deduction allows you to identify implicit knowledge and find contradictions in the ontology.

The types of rules for inferring additional knowledge are based on **Definition 6**. T*hese types can be used to model a subject domain with the help of a three-level ontology*. These rules allow to form the simple rules and their chains for the new links allocation.

The following relations are explicit:

thesaurus term ↔ information object,

information object ↔ semantic label,

thesaurus term ↔ classifier term.

Rule elements can be represented by variables of information resources and objects, constants-string and numeric expressions, predicates-linking attributes of various types, functional expressions-functions applied to individual arguments. The rules are set in the "if–then" form, for example:

– ***If*** *the description of the object contains the thesaurus term,* ***then*** *the "cloud" of keywords of the concept includes the keywords of the object*;

– ***If*** *the attribute value belongs to an object,* ***then*** *the concepts that describe that object can be grouped by the value of that attribute.*

We can word this rule as a requirement to create thematic "trends" based on the thesaurus and publications by year.

With the limited number of rule templates or meta-rules, their semantics are determined within a specific subject domain each time in their own way. The statements of the knowledge space ontology are used to determine relevant meta-rules and justify their use in the subject domain. They help to understand what can be further extracted from the library content (or the knowledge space).

10. MATHEMATICAL SUBJECT DOMAIN FEATURES

To support formula search, the concept of *Formula* was introduced. It allows you to store the original formula string obtained from the source. The string can be in the Content MathML, Presentation MathML, LATeX format [25]. If relevant, the number of formulas types representation in different notations is easily expanded. This concept of *Formula* is related to the objects, that make up the semantic library content, and the concepts of the thesaurus. Thus, it is possible to build a network of formula relations, both with the thesaurus concepts, and with various information objects of the system. Figure 9 shows such a network, with each node accessible from the *Formula* node.

Each formula can be supplemented with keywords. Keywords can be entered either by the system expert, or automatically, coming together with the formula from its source, as well as replenished with the keywords of related objects.

10.1. Search by Formulas

The search by formula consists of two logical parts-the search by formula itself and search by the keywords. Search by keywords is necessary to narrow down the candidates range. The search by formula should return formulas that are completely identical to the formula entered for the search or contain a part that is identical to the formula entered. The search algorithm can be divided into four phases:

1. Selection of candidate formulas. If relevant, convert formulas to MathML. At this stage, we get a list of formulas from the thesaurus that match the keyword search criteria.

2. Generating an internal representation for formulas. For each formula, we build an internal representation or use a pre-built internal representation.

3. Comparison of the desired formula with the candidate formulas for full or partial match (part of the candidate formula is equivalent to the desired formula).

4. Generating and displaying search results.

The selection of candidate formulas for keywords is as follows: the user enters keywords separated by a space. In case of at least one match of any keyword, the formula is included in the list of formulas to be compared with the desired formula.

After the query returns candidate formulas, we need to make sure that all of them have a representation in MathML format. If not, we need to convert formulas from LaTeX to MathML (formulas that don't have either LaTeX or MathML entries are not included in the search results). To convert formulas, you can use the MathToWeb library [26]. The conversion is performed in several threads to speed up the process. After that, you need to save the conversion results in the required field so that you can use them during the next search.

11. CONCLUSION

This paper features the approaches and methods for building a semantic library within the subject domain "Mathematics". The theoretical background of the work was based on ontologies in the construction of semantic libraries. This article describes the general ontology of the subject domain stepwise.

The proposed approach provides sufficient expressivity to be used when integrating different data sources. We identify the relations with subject domain thesaurus by means of the publication system, based on its title, annotation, and keywords. The Encyclopedia of Mathematics terms were used as semantic labels. With some degree of probability, such relating allow to identify articles from different sections of the subject domain and organize them into collections based on the MSC and UDC classifiers.

For other subject domains all the proposed methods of analysis and their modeling based on a three-level ontology are relevant as well.

FUNDING

This work was supported by budget topics of the Ministry of Science and Higher Education of the Russian Federation and particular by the Russian Foundation for Basic Research, project no. 20-07-00324.

REFERENCES

1. A. B. Antopol'skii, N. E. Kalenov, V. A. Serebryakov, and A. N. Sotnikov, "Common digital space of scientific knowledge," Vestn. Ross. Akad. Nauk **89**, 728–735 (2019).
2. A. B. Antopol'skii et al., "The principles of construction and structure of a unified digital space of scientific knowledge (UDSSK)," Nauch.-Tekh. Inform., Ser. 1, No. 4, 9–17 (2020).
3. N. I. Gubanov, N. N. Gubanov, and A. E. Volkov, "Criteria of the verity and scientific character of knowledge," Filos. Ob-vo **3** (80), 78–95 (2016).
4. V. V. Il'in and A. T. Kalinkin, *The Nature of Science: An Epistemological Analysis* (Vysshaya Shkola, Moscow, 1985; Progress Books, 1988).
5. O. M. Ataeva, N. E. Kalenov, and V. A. Serebryakov, "Ontological approach to the description of a common digital space of scientific knowledge," Ross. Tsifr. Bibl. Zh. **24** (1), 3–19 (2021).
6. A. B. Antopol'skii, O. M. Ataeva, and V. A. Serebryakov, "Environment of integration of data of scientific libraries, archives and museums LIBMETA," Inform. Resursy Ross. **5**, 8–12 (2012).
7. V. A. Serebryakov and O. M. Ataeva, "The basic concepts of a semantic libraries formal model and its integration process formalization," Program. Prod. Sist. **4**, 180–187 (2015).
8. O. M. Ataeva and V. A. Serebryakov, "The basic concepts of a semantic libraries formal model and its integration process formalization," Program. Prod. Sist. **11** (2), 85–100 (2017).
9. O. M. Ataeva, "LibMet semantic library information model," Program. Prod. Sist. **4**, 36–44 (2016).
10. O. M. Ataeva and V. A. Serebryakov, "LibMet digital semantic library ontology," Inform. Primen. **12**, 2–10 (2018).
11. A. M. Elizarov, A. V. Kirillovich, E. K. Lipachev, A. B. Zhizhchenko, and N. G. Zhil'tsov, "Mathematical knowledge ontologies and recommender systems for collections of documents in physics and mathematics," Dokl. Math. **93**, 231–233 (2016). doi 10.1134/S1064562416020174
12. A. A. Muromskii and N. P. Tuchkova, *About the Thesaurus for the Subject Area 'Ordinary Differential Equations'* (CCAS RAS, Moscow, 2004) [in Russian].
13. O. Ataeva, V. Serebryakov, and E. Sinelnikova, "Thesaurus and ontology building for semantic library based on mathematical encyclopedia," in *Proceedings of the 21st International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019),* CEUR Workshop **2523**, 148–157 (2019).
14. *Mathematical Encyclopedy,* Ed. by I. M. Vinogradov (Sov. Entsikl., Moscow, 1977) [in Russian].
15. Mathematics Subject Classification. http://msc2010.org/mediawiki/index.php?title=MSC2010. Accessed 2021.
16. Dbpedia. https://www.dbpedia.org/. Accessed 2021.
17. Math-Net.Ru. http://www.mathnet.ru. Accessed 2021.
18. Encyclopedia of Mathematics. https://www.encyclopediaofmath.org/index.php/Main_Page. Accessed 2021.

19. A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series*. *Special Functions* (Nauka, Moscow, 1983; Taylor Francis, London, 2002).
20. A. V. Vasil'ev, "Design pattern for enterprise Java applications built on responsive data models for scalability," Tr. MFTI **5** (4), 96–101 (2013).
21. A. M. Elizarov, A. V. Kirillovich, E. K. Lipachev, and O. A. Nevzorova, "Mathematical knowledge management: Ontological models and digital technology," CEUR Workshop Proc. **1752**, 44–50 (2016).
22. F. M. Bikmuratov, *Methods for Annotating Mathematical Texts* (Moscow, 2018) [in Russian].
23. Le Hoaj and A. F. Tuzovskii, "Semantic annotation of documents in digital libraries," Izv. Tomsk. Politekh. Univ. **322** (5), 157–164 (2013).
24. E. Oren, K. H. Moller, S. Scerri, S. Handschuh, and M. Sintek, What are Semantic Annotations? http://www.siegfriedhandschuh.net/pub/2006/whatissemannot2006.pdf. Accessed 2021.
25. A. M. Elizarov, E. K. Lipachev, and M. A. Malakhal'tsev, *Fundamentals of MathML. Submission of Mathematical Texts to the Internet* (Kazan, 2008) [in Russian].
26. Math To Web and User's Guide. http://www.mathtoweb.com/cgi-bin/mathtoweb_users_guide.pl. Accessed 2021.
27. M. M. Smirnov, *Equations of Mixed Type,* Vol. 51 of *American Translation of the Mathematical Monografs* (Am. Math. Soc., Philadelphia, 1988).
28. E. I. Moiseev, A. A. Muromskiy, and N. P. Tuchkova, *Information Retrieval Thesaurus in the Subject Area: Ordinary Differential Equations* (MAKS Press, Moscow, 2005) [in Russian].