Thesaurus and Ontology Building for Semantic Library Based on Mathematical Encyclopedia

Olga Ataeva[1], Vladimir Serebryakov[1], and Ekaterina Sinelnikova[2]

[1]Dorodnicyn Computing Centre, Federal Research Centre “Computer Science and Control” of the Russian Academy of Sciences, 40 Vavilov St., Moscow

[2]Moscow State University, 1 Leninskie Gory St., Moscow

Abstract. The paper focuses on the task of constructing the ontology of a thesaurus and its filling for the mathematics-related sources. The data of the Mathematical Encyclopedia, the Soviet encyclopedic edition, is used to fill the thesaurus of the subject domain and outlines it terminologically. The encyclopedia was preprocessed to structure its content and isolate semantic links automatically. The results were incorporated into the semantic e-library and used as a thesaurus for its subject domain.

Keywords: mathematical enciclopedia, ontology, thesaurus, digital library, semantic library

1 Introduction

This paper covers the task of constructing an ontology of a thesaurus and its filling of the library for the resources of the semantic library devoted to mathematics. The Mathematical Encyclopedia makes the thesaurus of the subject domain and outlines it terminologically. The Mathematical Encyclopedia is a five-volume Soviet encyclopedic issue that covers mathematical subjects. This fundamental illustrated edition on all main branches of mathematics comprising more than 6,000 articles was published in 1977–1985 [1]. Later the encyclopedia was digitalized. The e-version entries pose as an unstructured text with formulas in the form of images, lack any references to the related articles of the encyclopedia or other sources, and do not indicate the branch of mathematics. The listed drawbacks make the encyclopedia unfit for usage by Internet-users within the framework of an e-library.

To successfully integrate the data from the Mathematical Encyclopedia and make it available for e-library users, it is essential to ensure high level article structuring. Initially, the articles featured only text without any metadata such as references to the related articles of the encyclopedia, articles from other knowledge bases, or indication of a particular branch of mathematics. Such marking is important as the e-library user values not only the articles, but also options to navigate the library, to search relevant materials and related data.

The Mathematical Encyclopedia was translated into English in 1987 with about 2,000 new articles added. As of today, the e-version of the translated encyclopedia is supported by the international publisher Springer (Luxembourg) and is available on the Internet [2]. The Encyclopedia of Mathematics entries feature TEX-format formulas which can be machine-processed, and references to the related articles of the encyclopedia. Each article has a matching MSC (Mathematics Subject Classification) index [3] which is used for classification based on the branches of mathematics. Together, these metadata provide the user with a variety of options for searching relevant articles and studying related topics.

The data from the Encyclopedia of Mathematics is significant for adding new metadata to the Mathematical Encyclopedia and then making it available to the Internet-users as an electronic reference resource. This resource can be incorporated into a semantic e-library or provided with its own web-interface. In this paper, we have opted for LibMeta information system as such a library.

2 Overview

The overview of the related sources in Russian included online resources comprising articles from Mathematical Encyclopedia or any other mathematical knowledge.

Mathematics Library

This is a resource, available at www.MathemLib.ru, covers the knowledge accumulated throughout the Soviet period in a form of books published in the USSR and was updated with current news articles [4]. Moreover, this resource allows users to access articles from the Mathematical Encyclopedia listed alphabetically. It features the letters A-E, С, У from Russian alphabet and letters L, N, P, S from the Latin alphabet. The articles pose as a simple text with formulas in the form of images. It still lacks the references to other encyclopedia articles, articles from the library or any other sources.

Math-Net.Ru

It is the project of Steklov Mathematical Institute of Russian Academy of Science. The authors describe the project as a modern information system which provides Russian and foreign mathematicians with various options for searching information on mathematical life in Russia [5]. This information resource deals with mathematical journals and users have access to the full-version issues. Certain journals require paid subscription. The articles posted on Math-Net.Ru website also lack references to the related materials as they are available for downloading in PDF format.

World Digital Mathematical Library (WDML)

The researchers from Kazan Federal University cover the idea to create the mentioned resource in details in their papers [6]. The key purpose of WDML is to combine the digital versions of the entire corpus of mathematical academic literature within the distributed system of interrelated repositories, including both current resources and historical ones. Still, the ultimate goals of the project have not been attained.

The overview of the aforementioned and some other open mathematical resources in Russian has shown that there is an issue related to data structuring. The majority of the information resources do not allow users to build queries to the database, and observe relations between mathematical concepts, articles and authors.

3 Mathematical Encyclopedia Arrangement

To build information and reference resource based on the Mathematical Encyclopedia data, we have worked on including information on the branches of mathematics the articles belong to, placing cross-references between the articles, defining the article-related machine-readable formulas. The designed resource allows building queries to its contents as well as further integration with other databases within Linked Open Data space.

To attain the goal, we have addressed the data comprised in the English version of the encyclopedia – Encyclopedia of Mathematics. In particular, MSC section indexes and TEX-format formulas from the articles were employed. To employ them it is essential to match the Mathematical Encyclopedia articles with their translations from the Encyclopedia of Mathematics. Cross-references between the articles were generated by applying semantic annotation methods [7, 8]. A data model have been developed which helps build queries and ensure establishing links with other open sources.

Thus, the objective involved the following steps:

1. Develop the information resource data model based on the Mathematical Encyclopedia.
2. Match the Mathematical Encyclopedia articles with their translations in Ecyclopedia of Mathematics.
3. Add annotations, i.e. references to other encyclopedia entries, to the Mathematical Encyclopedia articles.
4. Attach MSC indexes to encyclopedia entries as in Encyclopedia of Mathematics.
5. Match articles with TEX-format formulas derived from the following Encyclopedia of Mathematics entries.
6. Correlate the entries with the list of similar articles.
7. Ensure building queries to the Mathematical Encyclopedia.
8. Ensure further integration of the developed resource with other sources.

3.1 Mathematical Encyclopedia Data Model

We have developed a data model that shows the relations between Mathematical Encyclopedia articles, terms, formulas, annotations and MSC sections. Fig. 1 illustrates the ontology classes and their relations.

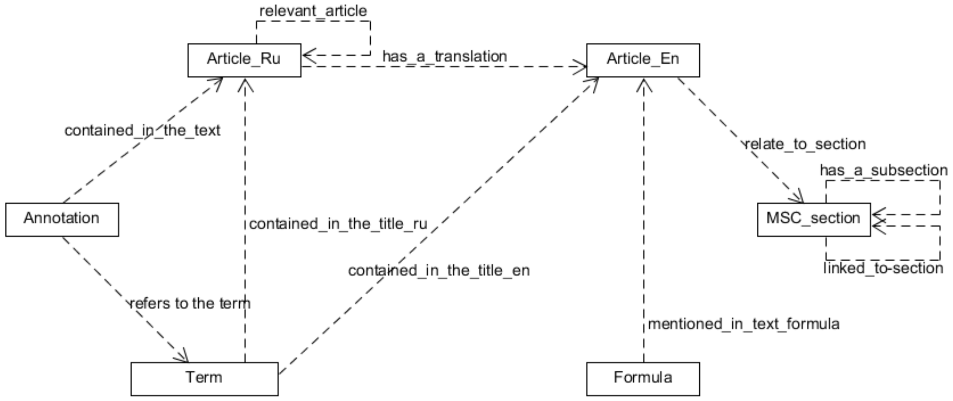

Fig. 1. Data Model

The ontology involves the following classes and relations:

1. ***Article_En* –** Encyclopedia of Mathematics article in English.
2. Object Properties: *relate_to_section* – indicates the MSC rubricator section which the article is related to.
3. ***Article_Ru*** – Mathematical Encyclopedia (ME) article in Russian.
4. Object Properties: *has_a_translation* – a reference to the item on the ***Article_En*** class which holds the translation of the Russian article; *relevant_article* – indicates the related article from ME.
5. ***Term*** – a mathematical concept which poses as a title to a certain article from ME or Encyclopedia of Mathematics.
6. Object Properties: *contained_in_the_title_ru* – indicates the ME article the title of which features the term; *contained_in_the_title_en* – indicates the Encyclopedia of Mathematics article the title of which features the term.
7. ***Annotation*** – an annotation found in the ME article text.
8. Object Properties: *contained_in_the_text Содержится_в_тексте_аннотация* – a reference to the ME article, the text of which features the annotation; *refers_to_the_term* – indicates the term being annotated.
9. Data Properties: *beginning_of_the_link* – the number of the word in the article which marks the beginning of the annotation*; end_of_the_link* – the number of the last word in the article included in the annotation.
10. ***Formula*** – a TEX-format formula included in the Encyclopedia of Mathematics article.
11. Object Properties: *mentioned_in_text_formula* – a reference to the Encyclopedia of Mathematics article which features the formula.
12. ***MSC_section*** – the section of the MSC mathematical rubricator.
13. Object Properties: *relate_to_section* – indicates the related MSC section, *has_a_section* indicates the close MSC section, *linked_to_section* – indicates the "parent" MSC section.

The developed model allows building queries to the Mathematical Encyclopedia. Let us consider some of these queries.

1. Find *ME Articles* that feature the *Term*.

The following is an example of a SPARQL query for the russian term “Multi-connected domain”:

```
SELECT  ?article_name
WHERE {
    ?annotation math_enc: contained_in_the_text> ?ru_article.
    ?ru_article rdfs:label ?article_name.
    ?annotation math_enc: refers_to_the_term> ?term.
    ?term rdfs:label ?label.
    filter contains(?label,"Многосвязная область")}
```

2. Find *ME* Articles that feature the *Formula*.

The following is an example of a SPARQL query for the formula $f\,(x_1, ..., x_n)$=0, presented in TeX notation

```
SELECT ?article_title
WHERE {
?ru_article math_enc: has_a_translation> ?en_article.
?ru_article rdfs:label ?article_title.
?formula math_enc: mentioned_in_text_formula> ?en_article.
?formula rdfs:isDefinedBy ?tex.
filter contains(?tex,"f(x_1,\\ldots,x_n)=0,\\tag{*}") }
```

3. Find *ME Articles* that relate to the *MSC Section*.

The following is an example of a SPARQL query for the *MSC – 60-XX* section – “Probability theory and stochastic processes”

```
SELECT  ?article_title
WHERE {
?ru_article math_enc: has_a_translation> ?en_article.
?ru_article rdfs:label ?article_title.
?en_article math_enc: relate_to_section> math_enc:60-XX>.
```

4. Show *ME Articles* that are relevant to the selected *ME article*.

The following is an example of a SPARQL query for a selection of articles relevant to the article “*Multi-connected domain*”

```
SELECT  ?article_title
WHERE {
math_enc:Article_Ru_Multi_connected_domain relevant_article
   ?see_also_article.
?see_also_article rdfs:label ?article_title.}
```

5. Show *ME Articles* that the annotations in the selected article reference to.

Below is an example of a SPARQL query for a selection of article links to which are found in the article “*Multi-connected domain*” in Russian version

```
SELECT  ?related_article_title
WHERE {
?annotation math_enc: contained_in_the_text math_enc: Arti-
   cle_Ru_Multi_connected_domain.
?annotation math_enc: refers_to_the_term ?termin.
?termin math_enc: contained_in_the_title_ru ?related_article.
?related_article rdfs:label ?related_article_title.}
```

6. Show *Formulas* related to the *Term*.

Below is an example of a SPARQL query for selecting formulas related to the term “*Pole*” in russian version

```
SELECT ?tex
WHERE {
?term  math_enc:Соде contained_in_the_title_ru ?ru_article.
?ru_article math_enc: has_a_translation ?en_article.
?formula math_enc: mentioned_in_text_formula ?en_article.
?formula rdfs:isDefinedBy ?tex.
?term rdfs:label ?label.
  filter contains(?label, «Полюс») }
```

7. Show *Formulas* related to the *MSC section*.

Below is an example of a SPARQL query for selecting formulas related to the *MSC section – 41-XX –* “*Approximations and expansions*”

```
SELECT ?tex
WHERE {
?formula math_enc: mentioned_in_text_formula ?en_article.
?en_article  math_enc: relate_to_section math_enc:41-XX.
?formula rdfs:isDefinedBy ?tex.}
```

Queries 1, 2 and 3 ensure the search through the Mathematical Encyclopedia articles based on such criteria as mathematical term, formula, and branch of mathematics. Query 4 and 5 show the relation between Mathematical Encyclopedia articles. Queries 6 and 7 is meant for observing the relations between formulas and mathematical terms, branches of mathematics.

Still, the listed queries do not explain the advantages of the ontological data model over other models concerning the current objective. One of the key distinctions of the data model from, for instance, the relational models is that the ontology can be processed by an inference engine which helps find out the relations between any two subjects within a model. Thus, the number of queries to the developed model is not limited to the aforementioned ones. Consider examples of queries that allow us to define new information templates based on existing information. Queries of this kind are also called rules.

8. Display formulas for the Russian article.

```
CONSTRUCT {?formula math_enc: mentioned_in_text_formula ?ru_article}
WHERE {
?ru_article math_enc: has_a_translation ?en_article.
?formula math_enc: mentioned_in_text_formula ?en_article}
```

9. Display MSC sections for Russian articles

```
CONSTRUCT {?ru_article math_enc: relate_to_section ?msc}
WHERE {
?ru_article math_enc: has_a_translation ?en_article.
?en_article math_enc: relate_to_section ?msc. }
```

Moreover, one of the objectives of the present paper is to ensure further linking of the developed data model to other knowledge sources within Linked Open Data. The structure of the relational models is rigidly fixed which makes the linking of two deferent models sophisticated and irresolvable in general. At the same time, ontologies can be linked relatively simply with the use of such OWL language properties as *owl:sameAs*, *owl:equivalentTo*.

Thus, the designed data model can serve as a frame for further information resource development based on the Mathematical Encyclopedia.

4 Thesaurus and Ontology Building for Semantic Library

Addressing the concept model described in the paper [9], as well as the ideas of Semantic Web and Linked Open Data, we have developed LibMeta *personal open semantic digital library* which supports the users' work with libraries' digital resources and collections within a particular scientific subject domain that is terminologically outlined by a thesaurus.

Apart from this, the key requirements to the system content, specifically *versatility, structure, adaptability* ensure support of the custom metadata repository for the objects as well as expanding information resource set. *Versatility* allows describing types of the system's resources and its objects regardless the subject domain or the users' scope of interest. The description *structure* supports relations between different external and internal resources relying on the LOD principles. The resource description *adaptability* allows adding new properties and links within the system development process and ensures user interface customization to reflect perspective changes. In fact, LibMeta makes the design of scientific knowledge space functional within the framework of a library.

The mathematical articles served as the subjects of the developed library. *Authors* and *Publications* were taken as examples of the resource types respectively. We have defined the set of attributes for each resource type within the minimum property set based on Dublin Core for publications and FOAF to describe the authors.

In fact, the concepts Authors and Publications serve as the items of the Information Resource class, which is defined as the basic unit of semantic library content. As each resource has a set of attributes, each of these items has its own assigned set of attributes that are described in the system. The set of attributes consists of the following elements: *title in the original language, title in Russian, surname, name, patronymic name, email address, date of birth, abstract, ID, author, occupation, publication type, place of birth, biography, description, additional title, language.*

4.1 Ontology of "Mathematical Encyclopedia" Thesaurus

The LibMeta thesaurus model is build to meet the standard, the ISO 25964 standard in particular [10]. This standard defines the thesaurus as a set of terms that are related by their respective links (relations).

The description of the mathematical encyclopedia in terms of the concepts of the basic version includes such concepts as *Thesaurus, Concept, Term, HierarchicalRelation, FamilyRelation.* The Mathematical Encyclopedia description in the terms of LibMeta ontology concepts can be additionally expanded. The attributes added are as follows: *formula, person, UDC code, MSC code, reference (to the English version of the concept).*

The attributes *reference* serves as the items of ***ThesaurusAttributeHref*** class, *formula*, *person* serve as the items of ***ThesaurusAttributeObject*** class. At the same time they make up the attribute set for the thesaurus, *UDC code, MSC code* are the items of ***ThesaurusAttributeTaxonomy*** class.

The concept structure of the Mathematical Encyclopedia lacks hierarchy as it is, still, due to the use of MSC codes related to the concepts we have managed to highlight related terms from certain branches of mathematics. We have derived the mentioned persons from the articles and linked concepts to the persons. Formulas were separately indexed and each concept, if possible, was matched to the set of respective formulas.

4.2 Mathematical Subject Area Features

To support the formula search within the sub-system, the concept *Formula* has been introduced and it helps store the original formula line from the resource it was derived from. The line might be featured in the following formats: Content MathML, Presentation MathML, LATeX. If needed, the number of formula representation types within different notations can be easily expanded. The concept *Formula* is related to *Information Objects* and *concepts* of the thesaurus. Thus, we can always build a network of formula relations with other system information objects and thesaurus concepts. Each formula can be updated with key words. Key words might be placed either by a system expert, or be added when they are derived automatically along with a formula from its resource, as well as be filled with key words from related objects.

5 Conclusion

The developed information resource allows studying Mathematical Encyclopedia articles, their relations, ensures their categorization regarding the branches of mathematics. The resource has the property of replenishment: the developed mechanisms can be applied to the new data to include it in the encyclopedia.

The further studies might address the development of the semantic article annotation. In particular, the researches might dwell on pseudonym support for the concepts. Another possible direction involves studying the correlation between MSC sections and other rubricators, for instance, UDC. If successful, it would be sufficient to link the ontology class that matches MSC sections, to a new class for UDC sections, so that we could categorize encyclopedia articles and related resources using a new rubricator.

The present study allowed us to employ a considerable amount of knowledge stored in the Mathematical Encyclopedia and then pass in on to the wide spectrum of amateur-users and experts in the mathematical field, which is particularly significant in the lack of open access to the similar resources.

The means of LibMeta system helped define the relations to the terms of Mathematical Encyclopedia for each publication based on its title, abstract and key words. This allowed us to carry out an additional thematic division of publications within the subject domain. To some degree, such linkage helped find the articles related to different branches of mathematics and arrange them in collections based on the thesaurus and placed MSC links. The study employed about 5,000 publications.

Acknowledgements

The study was supported by Russian Foundation for Basic Research, project
17-07-00214, 18-00-00-297.

References

1. Mathematical Encyclopedia. https://ru.wikipedia.org/wiki/ Математическая_энциклопедия (ru)
2. Encyclopedia of Mathematics. https://www.encyclopediaofmath.org/index.php/Main_Page (21.11.2018).
3. Mathematics Subject Classification. http://msc2010.org/mediawiki/index.php?title=MSC2010 (04.12.2018)
4. Math Library. http://www.mathemlib.ru (04.12.2018)
5. Math-Net.Ru. http://www.mathnet.ru
6. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., and Nevzorova, O.A.: Management of mathematical knowledge: ontological models and digital technologies. DAMDID/RCDL'2016, Ershovo, 11–14 oktyabrya 2016. P. 44–50 (ru).
7. Oren, E., Hinnerk Moller, K., Scerri, S., Handschuh, S., and Sintek, M.: What are Semantic Annotations? http://www.siegfriedhandschuh.net/pub/2006/whatissemannot2006.pdf (12.12.2018)
8. Le Hoaj, Tuzovskij, A.F.: Semantic annotation of documents in electronic libraries. News of Tomsk Polytechnic University **322** (5), 157–164 (2013) (ru).
9. Serebryakov, V.A. and Ataeva, O.M.: Information model of the open personal semantic library LibMeta. Proceedings of the XVIII Russian Scientific Conference "Scientific Service on the Internet". Novorossijsk, 19–24 September 2016. Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), 304-313 (2016) (ru).
10. ISO 25964 thesaurus schemas. http://www.niso.org/schemas/iso25964