Mathematical Modeling of the Processes of Interdisciplinary Collections Formation in the Digital Libraries Environment

N. Kalenov[1][0000-0001-5269-0988], I. Sobolevskaya[1][0000-0002-9461-3750], A. Sotnikov[1][0000-0002-0137-1255]

[1] Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences” (JSCC RAS — Branch of SRISA), 119334, Moscow, Leninsky av., 32 a, Russia
asotnikov@jscc.ru

Abstract. The task of forming a digital space of scientific knowledge (DSSK) is analyzed in the paper. The difference of this concept from the general concept of the information space is considered. The DSSK is represented as numerous different accessible objects verified by the world scientific community. The form of structured representations in the digital space is the semantic network, in which the fundamental principles of the organization are the fundamental objects and the subsequent construction of their hierarchy, in particular, according to the principle of inheritance. The classification of the objects that make up the content of the DSSK is introduced. The concept of a hierarchical relationship between an object is defined. The use of the concepts of set theory in the construction of DSSK allows you to divide by information into levels of detail. The concept of levels of the objects hierarchy of digital space of scientific knowledge is introduced. The definitions of objects of various levels are given. The principles of working with objects of each level are formulated too. It is shown that with the help of the hierarchical structure of information presentation in the digital library environment, a user collection can be formed in the central processing center. Constructing a hierarchy which is a section with a high degree of detail, allows you to increase the efficiency of information search in the space of knowledge and information analysis.

Keywords: Semantic Network, Information Space of Knowledge, Electronic Library, Levels of Detail, Hierarchy of Information Objects.

1 Introduction

Information is at the forefront of the many areas of our lives. IT and computing technologies penetration has expanded the possibilities for the capturing, analysis, disseminating, processing and use of scientific information.

Modern needments for professional information require the development of a knowledge space, which is a digital environment in which information resources and services from different fields of science, culture and education are integrated. A part of the general knowledge space is the digital space of scientific knowledge (DSSK), which differs from other components of the common space (in particular, such as Wikipedia) in that the information objects represented in the DSSK are verified by the world scientific community and are separated from information objects that are ideological, religious and other scientifically controversial character [1].

The flow of requests to the DSSK is often continuous, rapidly time-varying, not always predictable and unlimited in the form of the request. The software that processes such requests cannot afford to store and “review” the parameters of the request, which often requires a quick response in real time. Requirements for the accuracy of data retrieval in the DSSK (in contrast to the Search Engine) necessitate the development of special methods for request processing queries with a sufficiently accurate mapping of the query text to the metadata space describing certain objects of the DSSK. DSSK metadata, on the other hand, includes not only sets of keywords, but also more complex structures, for example, hierarchical classification systems.

The form of a structured representation of the digital knowledge space is a semantic network. The basic organization principle of this space is based on the classification system of objects and the subsequent construction of their hierarchy, in particular, according to the principle of inheritance: “macroeconomics” - section of the “economy”, “poetry collection” – publication, etc.

In accordance with this principle, objects are classified into a number of categories or classes based on their common features.

Most digital data collections are a diverse information network connecting objects of various types. For example, an electronic publication (the type of the object is “book”), in addition to plain text, contains additional information, such as the author of the publication (the type of the object is “person”), year of publication, publishing house, place of publication, etc. In turn, the object "person", in addition to a sequence of characters that specify a surname, is associated with person’s biography, a field of scientific interests ("subject matter"), etc. Thus, from the object “publication”, a connection can be established with another “object” (“author”), with the text of this publication, with the “subject of the object”, etc. In the general case, the DSSK should support various types of relationships between its elements – both within the same class of objects (in particular, a recursive hierarchical relationship) and between objects of different classes.

A significant number of studies have been devoted to the construction of thematic hierarchies, a hierarchy of concepts, object models, etc., which provide hierarchical organization of data at different levels of detail and have applications such as web search and viewing tasks [2, 3, 4].

In [5], the NetClus algorithm was described, which allows one to establish relationships between multi-type objects to create high-quality network clusters. The NetClus algorithm allows reordering attribute objects in each newly defined network cluster.

In this paper, we consider the DSSK in the aspect of theory of sets, which allows us to approach the issues of constructing space and working with it from a new point of view.

2 Configuration of DSSK

Let Ω – set containing all the elements of a digital scientific space located in some (possibly distributed, storage). Ω includes, in turn, two sets. The first of them (denoted by A) consists of digital images of real-world objects (digitized publications, archival documents, photographs, etc.) and objects created exclusively in the digital environment (electronic publications, 3D models, multimedia materials, etc.). All objects are being numbered in some way. Numbering should uniquely identify the object and provide the ability to retrieve it from storage.

The second set (denoted by B) includes metadata containing multidimensional characteristics of the objects of the first set, ensuring their selection by requests to DSSK and presentation to users.

The set A consists of elements a_i,, where i=*1*...N (N is the total number of objects reflected in Ω). These elements are objects of the following types:

- text files (recognized digitized printed or handwritten documents) or documents originally generated in electronic form;
- static images (unrecognized digitized documents, digitized or originally digitally generated photos);
- digital or digitized audio recordings;
- digital or digitized video/film materials;
- 3D models of various objects;
- multimedia installations (digital models of natural processes and technical devices, educational materials, virtual tours, etc.).

If the elements of the set A are being represented by a simple collection of pairs “object – its number”, then the set B in general, is a rather complex facet-hierarchical structure. Each of its elements is represented not only by a specific meaning and reference to an element of the set A (which is the case in traditional bibliographic information retrieval systems), but may include an indication of links with other elements. Thus, by elements of the set B we mean a structure that includes the semantic value of object characteristics, an indication to one or more elements of the set A corresponding this characteristic, and an indication ti relations with other structures, which are also elements of set B.

The constituent elements of set B can be indices of classification systems (such as State Classifier of Scientific and Technical Information, UDC, etc.) documents metadata such as individual characteristics of a person (surname and name, date and place of birth, etc.), names of events, their text descriptions, temporal and geographical characteristics of objects, etc.

In order to ensure the accuracy of searching for objects in the DSSK, set B must include a number of non-overlapping sets characterizing various aspects of information about the elements of the set A. Obviously, there can be myriad such partitions, but we will restrict ourselves to considering the “intuitive minimum”, but covering a wide range of characteristics of objects, a data set including classes such as “what (who), where, when”, supplemented by the “subject” class and formal characteristics specific to the DSSK, allocated to the subsets B_1 (types of objects listed above) and B_2 (conditions for providing users with various objects of the set A).

The subset B_1 of the set B ($B_1: B_1 \subset B$) consists of 6 elements, which are the characteristics of the elements of the metadata set, which we call the representation types of a digital object. Namely:

- b_{11} – text view with the ability to search for a fragment of text;
- b_{12} – static image;
- b_{13} – 3D object;
- b_{14} – audio document;
- b_{15} – video document;
- b_{16} – is a multimedia object.

The subset B_2 of the set B ($B_2: B_2 \subset B$) consists of elements that determine the conditions for the provision of a digital object to the user. The introduction of this subset is due to various legislative requirements for the public presentation of an object. Elements of the set B_2 will be called the conditions for the provision of the object. Namely:

- b_{21} – the object is in the free access;
- b_{24} – the object is in limited access, free of charge for a certain group of users (for example, a paid subscription to full-text scientific publications for employees of a certain institution) and inaccessible to other users;
- b_{23}– the object is in limited access free of charge for a certain group and commercial for other users (for example, a digital model of a museum exhibit may be available for free viewing to museum visitors, and remote viewing provides a certain fee.
- b_{24} – the object is commercially available, i.e. the user needs to pay access to this resource.

The subset B_3 of the set B ($B_3: B_3 \subset B$) contains the main characteristics of the object necessary for its identification during the search (“what” or “who”, “where”, “when”).

Note that the subsets B_1, B_2 and B_3 of the set B do not intersect each other.

The subset B_4 contains elements of the class “subjects”, it can have a rather complex structure containing indices and names of elements of various classification systems (strictly hierarchical type State Rubricator of Scientific and Technical Information [7], facetted type UDC [8, 9], etc.), keywords and terms, thesauruses, etc.

2.1 Examples of Subsets B_3

As an element of the set B_3 belonging to the class “what”, to act as an obligatory element is the name of a specific object, which can be supplemented by elements that specify the type of object within a given view, such as a book, article, archive document, etc., as well as unstructured explanations containing this or that information about the object. For example, a collection of photos of Moscow of the 30s can be supplemented by a detailed article on the architecture of the city of that time, presented in the form of hypertext.

As an element of the “where” class of the set B_3, various implementations can be made that are related directly to the geographical location of the object (for example, for a person the place of birth, for a museum object - the place of its initial discovery, for an event - the country or city where it occurred, etc.). Elements that are belonged to class "where" may be represent as organizations described, in turn, by its metadata (for example, the person’s place of work, the place of storage of a museum item, a publishing house for a printed document, etc.). For example, the collection of herbaria A.N. Petunnikova [http://www.gbmt.ru/ru/about/fund/fondovaya-kollektsiya-gerbariy/, http://e-heritage.ru/ras/view/person/general.html?id=49901007], contains information about geographic location of the collection elements.

The print materials year of publication, or year of the person birth, etc. may act as an element of the “when” class of the set B_3.

3 The Collections Generation of the Set A Elements

If the user’s requests to the DSSK include only elements of the set B, then the task of selecting and presenting documents is reduced to the formation of the sampling conditions, consisting of query terms connected each other by Boolean operators. Data search in space B is carried out by comparing its elements with query elements (let's call it linear search). The result of the search is the addresses of the corresponding elements of the set A, by which these elements are retrieved from the store and provided to the user in accordance with the conditions reflected in the set B_2.

However, in practice, often the user needs to create a collection of elements of the set A for requests that do not explicitly formulate in terms of the set B. For example, it is necessary to form a user's collection of the "Silver Age" poets works reflected in the digital library (DL) of 20th century publications. During the formation of the DL elements the “Silver Age” was not indicated as a temporary characteristic. It is also not present in any of the classification systems that could be used in the database creation. In accordance with a query in terms of “Silver Age poets” its result will be an empty subset of the elements of the set A. At the same time it is obvious that among the elements of the set A there are objects that meet the requirements for this collection. To detect them it is necessary to construct a mapping of the user query to the set B and then implement a linear search for the query, including the corresponding elements of the set B.

Note that the properties of objects of the space Ω are transferred to all objects of the subsets of this set, which avoids a significant part of the duplication of information [10, 11].

4 Building a Hierarchy of Digital Objects Representation in a Digital Library Environment

Let the set F be the set of “the object’s characteristics”. The characteristic of the object (f_s, where $s = 1 \dots \infty$) is understood as a given parameter, according to which the objects of the set A will be combined into a user collection. For example, some research area or object type (mineral, silver age poets, minerals mentioned in silver age poetry, etc.) could be such parameter. I.e. $F = \bigcup_s^\infty f_s$.

Applying some mapping f_s to the elements of the sets A and B, we can obtain the so-called collection that is the subset of one or more types objects integrated by a given parameters: $f_s(B(A), A)$.

Applying the map f_s to the set B, we obtain the subset $B(f_s)$ by which we can select the objects corresponding to it from the set A. We call the class of these objects the user collection $G(f_s)$. The collection of subsets A corresponding to some common characteristic $B(f)$ is called the class set $G(f)$. The concepts and definitions described above make it possible to build a hierarchy of the representation of digital library environment objects which allows to formalize of a general approach to the formation of user collections conversely.

The hierarchy levels of user collections are determined by the structure of user-defined object mapping properties to subsets B. In general:

$$B(f_s) = B_1(f) \cup B_3(f) \cup B_4(f)$$

The first level of the hierarchy (narrow-focus collections) includes collections for which the subset $B_3(f)$ is not empty; the second level (theme-specific collections) includes collections for which $B_3(f)$ is empty and $B_1(f)$ and $B_4(f)$ are not empty; the third level (thematic collections) includes collections for which $B_1(f)$ and $B_3(f)$ are empty and not empty $B_4(f)$.

5 Conclusion

The creating of such a hierarchy allows to optimize the process of formation and maintenance of information funds of digital libraries and also allows the user to choose from the whole set of interconnected resources of the digital library those information objects that are united by one or more features [11–16].

Using the hierarchical structure of the information representation of objects in the environment of the digital library, a user collection can be formed in the DSSK.

Thus the task of collections creation according to a prescribed criterion is reduced to the following steps (Figure 1):

1. A correspondence analysis of the elements of this collection attribute to the elements of the set B;
2. Separation the collection attributes into two subsets. The first subset contains the elements of the set B in explicit form (for example, the type of an object, typical objects, etc.). The second subset doesn't contains the elements of the set B in explicit form;
3. The implementation of the algorithm for mapping the characteristics of the collection to the set B.

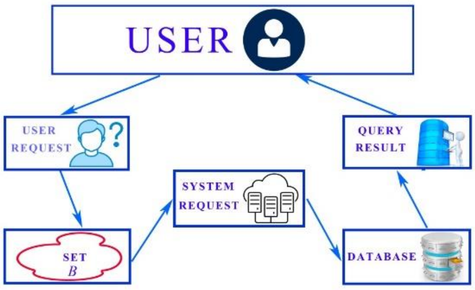

Fig. 1. Users collections creation

As an example of the implementation of the algorithm for mapping the collection to a set of metadata, let us consider the formation of materials collection related to the "Silver Age" poets in the environment of the Digital Library "Scientific Heritage of Russia" (DL SHR). Such collection should include information about the authors and all documents related to them (including bibliography and full texts of their works).

At the first stage, the years of publication are determined. I.e. the parameter of the request "Silver Age" is translated into the "language" of metadata. After that, a selection is made of all objects included in the DL SHR for a given span of time. Next, from the found data array, persons are selected that correspond to such metadata elements as the "author", which, in turn, are associated with publications that have a "publication type" metadata element with the value "poetry". As a result, we get a list of the names of the poets of the Silver Age.

Then, of all the objects obtained in the first stage, all materials are selected according to the "author" parameter.

Thus, a collection of all DL SHR objects associated with the "Silver Age" poets will be obtained (including archival documents, photo documents, etc.).

The research is carried out by JSCC RAS – branch of SRISA within supported by Russian Foundation for Basic Research (projects 18-07-00893 and 18-00-00372).

References

1. Antopol'skij, A.B., Kalenov, N.E., Serebryakov, V.A., Sotnikov, N.A.: Tochka zreniya o edinom cifrovom prostranstve nauchnyh znanij. Vestnik Rossijskoj akademii nauk 89(7), pp. 728–735 (2019).
2. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intell Agent Syst. 1(3, 4), pp. 219–234 (2003).
3. Sun, Y., Yu, Y., Han, J.: Ranking-based clustering of heterogeneous information networks with star network schema. KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 797–806 (2009).
4. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: a look back and into the future. ACM Computing Surveys (CSUR) 44(4), Article 20 (2012).

5. Wang, C., Liu, J., Desai, N., Danilevsky, M., Han, J.: Constructing topical hierarchies in heterogeneous information networks. Knowledge and Information Systems 44(3), pp. 529–558 (2015).
6. Kalenov N.E., Sobolevskaya I.N., Sotnikov A.N. Ierarhicheskie urovni predstavleniya informacionnyh ob"ektov v srede elektronnyh bibliotek. Informaciya i innovacii. Vol. 13. Iss. 2. pp. 25–31 (2018).
7. Kalenov, N., Sobolevskaya, I., Sotnikov, A.: Digital museum collections and representation of objects of natural history museum storage in the Scientific Heritage of Russia Digital Library. Scientific and technical information Ser. 1, 10, pp. 33–38 (2016).
8. Antopol'skij, A.B., Beloozerov, V.N., Markarova, T.S., Dmitrieva, E.Y.: Ustanovlenie sootvetstvij rubrik GRNTI rubrikam drugih sistem klassifikacii nauchnoj i tekhnicheskoj informacii. Scientific and technical information Ser. 1, 3, pp. 3–18 (2015).
9. Astahova, T.S.: Problemy otrazheniya sovremennogo nauchnogo znaniya v klassifikacionnyh sistemah: novoe v UDK. Sbornik trudov konferencii «Perspektivnye napravleniya nauchnyh issledovanij i kriticheskie tekhnologii v klassifikacionnyh sistemah». VINITI RAN, Moskva, 25-27 oktyabrya, pp. 32–35, Moscow (2017).
10. Aleksandrov, P.S.: Vvedenie v teoriyu mnozhestv i obshchuyu topologiyu. Nauka, Moscow (1977).
11. Steffen, L., Manat, M., Frank, S.: Reductions between types of numberings. Annals of Pure and Applied Logic 170 (12), 102716 (2019).
12. Antopolsky, A., Atayeva, O., Serebryakov, V.: Environment of integration of data of scientific libraries, archives, and museums “LibMeta”. Information resources of Russia Vol. 5(129), pp. 8–12 (2012).
13. Kalenov, N., Sobolevskaya, I., Sotnikov, A.: On the interaction of the Scientific Heritage of Russia Digital Library with natural history museums. Information resources of Russia 148, pp. 2–6 (2015).
14. Ivanov, V.M., Strelkov, S.V., Kholina, A.A., Avtyushenko, A.L.: Virtual reconstructions in multimedia exhibitions of objects of cultural heritage. Virtual archaeology collection Hermitage, pp. 41–49 (2015), http://www.virtualarchaeology.ru/pdf/281_va_book2015.pdf, last accessed 2019/11/12.
15. Barutkina, L.P.: Multimedia in a modern museum exhibition. Bulletin of St. Petersburg State University of Culture and Arts. SPbSUCA, pp. 106–108 (2011).
16. Vassileva, S., Kovatcheva, E.: The innovative model for interactivity in Bulgarian museums. In: 10th Annual International Conference of Education, Research and Innovation (ICERI), CERI Proceedings, pp. 5407–5412 (2017).
17. Maggio, A., Kuffer, J., Lazzari, M.: Advances and trends in bibliographic research: Examples of new technological applications for the cataloging of the georeferenced library heritage. Journal of Librarianship and Information Science 49(3), pp. 299–312 (2017).
18. Frandsen, T.F., Tibyampansha, D., Ibrahim, G.R., von Isenburg, M.: Library training to promote electronic resource usage: A case study in information literacy assessment. Information and Learning Science 118(11–12), pp. 618–628 (2017).
19. Shahzad, F., Alwosaibi, F.M. Development of an e-Library Web application. IMSCI. In: 11th International Multi-Conference on Society; Orlando; the United States, Cybernetics and Informatics, Proceedings, pp. 153–158 (2017).
20. Mi, X.Y., Pollock, B.M.: Metadata Schema to Facilitate Linked Data for 3D Digital Models of Cultural Heritage Collections: A University of South Florida Libraries Case Study. Cataloging & Classification Quarterly. 56(2–3), pp. 273–286 (2018).