Digital Library Metadata Factories

Alexander Elizarov[1, 2][0000-0003-2546-6897] and Evgeny Lipachev[1] [0000-0001-7789-2332]

[1] Kazan (Volga Region) Federal University, Russia
[2] Kazan Branch of the Interdepartmental Supercomputer Center of the Russian Academy of Sciences, Russia
amelizarov@gmail.com, elipachev@gmail.com

Abstract. As you know, today the digital economy is understood as an economy based on the processes of production and use of digital technologies. Currently, these processes are largely implemented on the basis of digital platforms organized in various subject areas and fields of activity. Such platforms have their own sets of services and allow solving various sets of tasks for the development and use of digital technologies.
This article discusses the development and use of digital technology in scientific activities based on digital platforms. Such platforms have their own sets of services and allow solving various sets of tasks for the development and use of digital technologies. We have indicated the importance and role of digital libraries in the formation of digital platforms, and also analyzed the problems of ensuring the integration (connectivity) of the extracted information.
The concept of a metadata factory is presented by us as a system of interconnected software tools designed to create, process, store and manage metadata of digital library objects. Such metadata factories make it possible to integrate created electronic collections into digital scientific libraries that will combine these collections.
We have solved a number of problems associated with the construction of the metadata factory of the digital mathematical library named after Lobachevsky-DML. We suggest using this implemented metadata factory as an element of the ecosystem of any scientific digital library.

Keywords: Digital Science Platform, Digital Science Library, Digital Library Ecosystem, Metadata, Metadata Generation, Metadata Extraction, Metadata Normalization, Metadata Factory, Digital Mathematical Library Lobachevskii-DML.

Introduction

As you know, at present, the digital economy is understood as an economy based on the processes of production and use of digital technologies (see, for example, [1]). Today, these processes are largely implemented on the basis of digital platforms organized in various subject areas and fields of activity. Such platforms have their own sets of services and allow you to solve a variety of problem sets for the development and use of digital technologies. The very concept of a digital platform and some features

of the current stage of their development are analyzed in [2]. In [3], estimates are given of the current level of development of digital platforms in Russia.
These levels include the following four main components: the definition of multilateral digital platforms, development factors for digital platforms, business models, and competition dynamics. Note that the formation of digital platforms for research and development is provided for by the Digital Economy of the Russian Federation program [4, Ch. 1]. This program emphasizes the need to create digital platforms for basic and applied research, research and development. It is research and development that precisely constitutes the main lines of activity in the field of research and development. As a rule, each created digital platform has a specific organization that acts as a platform operator and forms its own ecosystem around itself.

Another area of development and use of digital technologies in scientific activity provides for the organization of access to the latest scientific results, in particular, scientific publications and scientometric information about them, using modern digital technologies. Historically, this direction is associated with the formation of digital (or electronic) libraries in the world, including scientific ones. Their active development began at the end of the twentieth century (see, for example, [5–7]). In general, digital (electronic) libraries of any orientation (not only scientific) mean models of complex information systems that serve as the basis for creating universal distributed knowledge storages and are equipped with navigation and search tools in the collections of heterogeneous electronic documents included in these storages.

1 Digital Scientific Libraries

Currently, digital scientific libraries exist in all developed countries of the world [7, 8].

1.1 Russian Digital Scientific Libraries

We only note the most famous Russian digital scientific libraries that provide not only access to scientific content, but also services for working with it.

- Socionet (https://socionet.ru, year of organization – 2000). This is a digital library that ensured Russia's participation in the development of an international online scientific and educational infrastructure (initially – in the field of social sciences, now – in all scientific disciplines);
- eLibrary (https://elibrary.ru/, year of organization – 2005). This is the largest Russian digital scientific library integrated with the Russian Science Citation Index (RSCI);
- Cyberleninka (https://cyberleninka.ru, year of organization – 2012). This is a digital scientific library, which is based on the concept of open science (Open Science) and is one of the five largest open scientific archives in the world;
- MathNet.RU (http://www.mathnet.ru/, year of organization – 2006). This is an all-Russian mathematical portal. It contains archives of leading Russian mathematical journals, collections of video lectures, navigation and search tools, as well as an information system for managing editorial processes [9].

In Russia in addition to those listed a large number of various digital libraries have been created related to modern publishing and scientometric services.

Examples of the latters are Mendeley (https://www.mendeley.com, year of organization – 2008) is a tool for managing a personal scientific library and effective collaborative scientific work; ISTINA (https://istina.msu.ru/, year of organization – 2014) is a digital platform for collecting, organizing, storing and analyzing scientometric information for the preparation and adoption of managerial decisions.

1.2 International Digital Science Libraries

The largest international digital scientific libraries are scientometric databases:

- Web of Science (until 2014 – Web of Knowledge) (https://clarivate.com/webofsciencegroup/solutions/web-of-science/). It arose in 1961 as a product of the American company ISI (Institute for Scientific Information), later belonged to the Thomson Reuters media corporation and became digital, since 2016 it belongs to Clarivate Analytics; its main product is Web of Science Core Collection;
- Scopus (https://www.scopus.com/, founded in 2004). This is the largest citation database for peer-reviewed scientific literature.

These digital libraries (as well as a large number of others) play a huge role in accelerating the circulation of existing knowledge and access to it. But without the Internet, which today has become a comprehensive integrated information environment, extracting information from various kinds of information sources (databases), which are a variety of digital libraries, would be impossible. At the same time, a number of serious problems arise in ensuring the integration (connectivity) of the extracted information. From this point of view, the narrowing of the entire space of available information makes it possible to more accurately specify information and, therefore, provide better access to and use of it. Such a narrowing is provided in the framework of specialized digital scientific libraries, which are organized in specific subject areas. For example, mathematical digital libraries have reached a high level of organization. The history of their origin and development is presented, for example, in [8, 10]. Digital libraries such as MathNet [9], Numdam [11], dml-cz [12] provide services that take into account the peculiarities of mathematical content. Within the framework of the project The European Digital Mathematics Library (EuDML, https://initiative.eudml.org/), methods for integrating European digital mathematical collections are being developed.

Thus, at present, in the field of science and scientific research, on the one hand, a significant number of different digital scientific libraries have been formed. They implement a wide range of search services. Each digital library has its own ecosystem. On the other hand, there are currently no examples of digital scientific platforms that are created and in accordance with the basic definitions [2, 3] successfully implement their own functions and user interaction services, as well as their own business models. We believe that digital scientific libraries can serve as the basis for building such digital platforms. At the same time, existing (not all) ecosystems must be improved. In the absence of such ecosystems, they simply need to be created. Below we discuss one area of such improvement.

2 Metadata and Navigation in the Scientific Information Space

Currently, digital scientific libraries exist in all developed countries of the world [7, 8]. It is well known that today navigation in the information space is largely provided by the availability and completeness of the set of metadata (data about data) of documents presented on the network (for example, [13–16]). There are currently quite a few different metadata standards. These standards should provide opportunities for interoperability with the external environment, identification and integration of information, its search in a distributed environment. Metadata should be open and extensible, oriented to modern semantic and digital technologies. But even with such standards, it is very difficult to really provide the necessary metadata properties for various documents. It is much easier to standardize metadata in relation to a specific subject area and on the basis of those digital libraries that are created in this area. An example is the field of mathematical and computer sciences, where a significant number of digital libraries have been created that perform various functions of integrating mathematical knowledge. Features of the presentation of document metadata in various digital mathematical libraries are described in [10, 11].

2.1 Digital Library Metadata Factory

We believe that a metadata factory should become an essential element of the ecosystem of any digital library. We use the term "metadata factory of digital library" in the following sense: a metadata factory is a system of interconnected software tools aimed at creating, processing, storing and managing metadata of digital library objects and allowing integrating created electronic collections into aggregating digital scientific libraries. The use of these tools, mainly in the automatic mode, will ensure the performance of operations such as selecting objects and their relationships, extracting metadata from various sources and specific documents, checking, refining, improving, normalizing in various formats, and matching metadata (using manual editing or automatic agents), as well as storing and linking metadata to external databases. In the case of the digital mathematical library, a number of specialized ones are added to the listed tools. For example, this is a conversion to the MathML format, markup of mathematical formulas and organization of a search on them [11, 17–19].

We indicate the main tasks that must be solved within the framework of the digital library metadata factory.

When working with digital libraries, one of the important tasks is the automated integration of the repositories of relevant documents with other information systems. Such a process is based on a model of aggregation and dissemination of metadata. The OAI Protocol for Metadata Harvesting model (http://www.openarchives.org/OAI/openarchivesprotocol.html, hereinafter OAI-PMH) is supported by most systems designed to store information resources. To organize work with OAI-PMH it is necessary to use a digital storage support system. The most famous of these are DSpace, Eprints, Fedora, and Greenstone. Some libraries have specialized methods for harvesting metadata from other repositories. In this case, it is necessary that the data providers have tools and services that allow the dissemination of metadata.

To organize the interaction of services both within the digital library and with external libraries and databases, it is necessary to take into account the metadata formats that are used in them. Even in one digital library, software tools work with multiple metadata formats. This is due both to the features of the formation of digital content, and to the requirements of aggregating digital libraries and scientometric databases. We mention only the most common metadata formats that you have to deal with when organizing the interaction of services in digital libraries (their full descriptions are available on the Internet).

First of all, this is the Dublin Core format and its extensions, the MARC cataloging format, RIS (Research Information Systems) bibliographic link formats, AMSBib, and the Russian Science Citation Index (RSCI) XML format.

Separately, we note the XML schemes of the Journal Archiving and Interchange Tag Suite (NISO JATS), which are designed to meta-describe articles in scientific journals [20, 21]. The significance of these schemes for digital mathematical libraries is determined by the fact that the mandatory and fundamental metadata sets for The European Digital Mathematics Library (EuDML) [22] are based on the NISO JATS v.1.0 scheme.

Many of the tasks mentioned above were solved by us when constructing the metadata factory of the digital mathematical library Lobachevskii-DML (https://lobachevskii-dml.ru/). As in the case of any digital scientific library, the formation of the Lobachevsky-DML library and the corresponding metadata factory required the use of technological solutions to manage scientific content, both previously created and those that were newly developed by us.

The formation of metadata of digital mathematical collections in the metadata factory of the digital library Lobachevskii-DML is carried out in several stages.

At the stage of preprocessing, the collection of documents is processed by software tools in order to bring it to a form suitable for further automatic processing [10]. For this, clustering of documents by stylistic similarity is performed, and then the stylized constructions used in the document are reduced to the template form. For example, .tex collections may contain documents that use not only the `\title{}` command, but also `\tit{}`, `\ArticleNAME{}` and others to design the article title. Lists of authors, keywords, codes of subject classifiers and other information necessary for the formation of a set of metadata are recorded using tex-commands, which differ significantly in the style files of various journals.

Moreover, the title of the article, the list of authors, keywords and other blocks necessary for inclusion in the metadata may not be executed by teams. In this case, they simply differ in font selection, for example, `{\bf Paper Title}`. In these cases, an attempt is made to automatically find such blocks by location in the text and font design. This approach applies to most collections of documents created in office formats [23]. In [24–26], algorithms for extracting metadata from scientific articles are presented. These algorithms are based on the study of the structure of documents and ontologies for describing the structure of documents.

At the preprocessing stage, part of the files cannot be processed automatically. For example, this situation occurs when processing a tex-document, when there is no style file or a file with author macro definitions that are referenced in the processed document. From such files a set is formed, which is adjusted by semi-automatic tools or

manually. After that, preprocessing is repeated. Files rejected at this stage several times are manually corrected.

The next step is the formation of a set of basic metadata. Strings with the title of the article and the list of authors are extracted from the documents. Next, the search is carried out and the text of codes of subject classification, a block of keywords, affiliation of authors and abstracts to the article, if they are given in the document, are extracted. Also, the metadata of each document includes its URL-link in the digital library collection. The generated metadata is stored in the xml-file in accordance with the DTD-rules and XML-schemes, which are installed in the digital library.

At the stage of improvement and refinement of metadata, software tools are used, with which simple spelling errors and typos are corrected in the title of the article, list of authors and keywords. Metadata is being improved, in particular, transliteration of article titles, addition of abbreviations with full titles ("SPb" –"Saint Petersburg", "LJM" – "Lobachevskii J. Math.", "Lobachevskii Journal of Mathematics").

URLs are checked using existing refinement services. Then, the metadata includes the date the web resource was accessed. Formula fragments in article titles and annotations are converted to MathML code.

Not all metadata can be obtained by searching for the corresponding blocks in the document and then extracting it from the text. Keywords, classifier codes, and other data are determined only as the result of textual and semantic analysis of the document [18, 24]. In the metadata factory, these operations are performed at the stage of generating additional metadata. As part of the project to create a digital mathematical library, we have developed tools to automate a number of operations at this stage.

2.2 Normalization of Metadata

We use the term normalization [10] to refer to the methods for generating and converting document metadata in accordance with the rules and XML schemes of digital libraries and scientometric databases.

One of the functions of the metadata factory is the normalization of metadata in accordance with the formats of other aggregating libraries. For example, the OAI-PMH protocol requires the inclusion of a metadata set in the resource description in the oai_dc notation, which is based on Dublin Core.

In the metadata factory of the digital mathematical library Lobachevskii-DML, a method has been developed for normalizing metadata into the format of the Russian Science Citation Index [27]. This method is implemented as a plugin of the journal platform Open Journal System and is used to generate metadata of the digital journal "Russian Digital Libraries Journal" (https://elbib.ru/). Methods have also been created to normalize the metadata of the collection of articles of this journal into formats of the bibliographic database on computer sciences "dblp computer science bibliography" (DBLP, https://dblp.uni-trier.de/) [28]. Methods have been created for the formation of mandatory and fundamental sets of metadata using XML schemes of the EuDML European Mathematical Library (https://initiative.eudml.org/) [10, 28].

Tools of the metadata factory of the Lobachevskii-DML library, already implemented by us, are described in detail in [10].

Conclusion

So, as a result of the development of the metadata factory of the digital scientific library Lobachevskii-DML:

- a system of services has been proposed for the automated generation of metadata of electronic mathematical collections;
- developed an xml-language for the presentation of metadata, based on the Journal Archiving and Interchange Tag Suite (NISO JATS);
- software tools have been created to normalize metadata of electronic collections of scientific documents in formats developed by international organizations - aggregators of resources in mathematics and Computer Science;
- an algorithm has been developed for converting metadata to the oai_dc format and generating the archive structure for import into DSpace digital storage;
- methods for integrating electronic mathematical collections of Kazan University into domestic and foreign digital mathematical libraries have been proposed and implemented.

The metadata factory model presented above was implemented in a specific digital mathematical library. Naturally, this model takes into account the specifics of the processed content. At the same time, it can be used as an element of the ecosystem of any scientific digital library.

Acknowledgement. This work was partially supported by the Russian Foundation for Basic Research under the project No. 18-29-03086, the Russian Foundation for Basic Research and the Government of the Republic of Tatarstan under the project No. 18-47-160012 and the Development program of the Regional Scientific and Educational Mathematical Center Volga Federal District, agreement number No. 075-02-2020-1478/1. This article also contains the results obtained in the framework of the project “Monitoring and standardization of the development and use of technologies for storing and analyzing big data in the digital economy of the Russian Federation”, carried out as part of the Program of Competence Center of the National Technological Initiative “Center for Storage and Analysis of Large Data” supported by the Ministry of Science and Higher Education of the Russian Federation under the Treaty of the Lomonosov Moscow State University with the Project Support Fund of the National Technology Initiative dated 08/15/2019 No. 7/1251/2019.

References

1. Ershova, T.V.: The Conceptualization of the Subject Area ‘Digital Economy’ as the Basis for the Development of Its Terminological Framework. Information society (6), 34–41 (2019).
2. Ershova, T.V., Hohlov, Yu.E.: Digital Research & Development Platforms. Information society (6), 17–24 (2017).
3. Eferin, Ya.Yu., Rossoto, K.M., Hohlov, Yu.E.: Digital Platforms in Russia: Competition between National and Foreign Multisided Platforms Stimulates Growth and Innovation. Information society (1-2), 16–34 (2019).

4. Program «Digital economy of the Russian Federation». 28 July 2017. No. 1632, http://government.ru/docs/28653/, last accessed 2020/7/10. [in Russian].
5. Arms, W.Y.: Digital libraries. Cambridge; London (2000).
6. Antopol'skij, A.B., Majstrovich, T.V.: E-Libraries: Principles of Creation. Moscow, Liberiya-Bibinform (2007). [in Russian].
7. Xie, I., Matusiak, K.K.: Discover Digital Libraries: Theory and Practice. Elsevier Inc. (2016).
8. Elizarov, A.M, Lipachev, E.K, Zuev, D.S.: Digital mathematical libraries: Overview of implementations and content management services. CEUR Workshop Proceedings 2022, 317–325 (2017).
9. Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., Zhizhchenko, A.B.: MathNet.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics. LNCS 7961, 344–348 (2013). https://doi.org/10.1007/978-3-642-39320-4_26.
10. Gafurova, P.O., Elizarov, A.M., Lipachev, E.K.: Basic Services of Factory Metadata Digital Mathematical Library Lobachevskii-dml. Russian Digital Libraries Journal 23 (3), 336–381 (2020).
11. Bouche, T., Labbe, O.: The New Numdam Platform. CICM 2017: Intelligent Computer Mathematics, 70–82 (2017).
12. Bartošek, M., Rákosník, J.: DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library. Notices of the AMS 60 (8), 1028–1033 (2013). http://dx.doi.org/10.1090/noti1031.
13. Sicilia, M.-A. (Ed.): Handbook of Metadata, Semantics and Ontologies. World Scientific Publishing Co. Pte. Ltd. (2014).
14. Alemu, G., Stevens, B.: An Emergent Theory of Digital Library Metadata. Enrich then Filter. Chandos Publishing is an imprint of Elsevier (2015).
15. Gartner, R.: Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer International Publishing Switzerland (2016).
16. Kogalovsky, M.R.: Metadata in Computer Systems. Programming and Computer Software. 39 (4), 182–193 (2013).
17. Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O.: Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management. Communications in Computer and Information Science, Springer 706, 33–46 (2017). doi:10.1007/978-3-319-57135-5_3.
18. Elizarov, A.M., Lipachev, E.K.: Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University. CEUR Workshop Proceedings, 2022, 326–333 (2017).
19. Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O.: Semantic Formula Search in Digital Mathematical Libraries. In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE, pp. 39–43 (2017). https://doi.org/10.1109/RPC.2017.8168063.
20. Journal Publishing Tag Library NISO JATS, version 1.3d1 (ANSI/NISO Z39.96-2019). October 2019, https://jats.nlm.nih.gov/archiving/1.3d1/, last accessed 2020/7/10.
21. Jost, M., Bouche, T., Goutorbe, C., Jorda, J.P.: D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010, http://www.mathdoc.fr/publis/d3.2-v1.6.pdf, last accessed 2020/7/10.
22. EuDML metadata schema specification (v2.0–final),https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2020/7/10.
23. Standard ECMA-376: Office Open XML File Formats, https://www.ecma-international.org/publications/standards/Ecma-376.htm, last accessed 2020/7/10.

24. Elizarov, A.M., Lipachev, E.K., Khaydarov, S.M.: Automated System of Services for Processing of Large Collections of Scientific Documents. CEUR Workshop Proceedings. 1752, 58–64 (2016).
25. Elizarov, A.M., Khaydarov, Sh.M., Lipachev, E.K.: Scientific Documents Ontologies for Semantic Representation of Digital Libraries. In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE, pp. 1–5. https://doi.org/10.1109/RPC.2017.8168064.
26. Tkaczyk, D., Tarnawski, B., Bolikowski Ł.: Structured Affiliations Extraction from Scientific Literature. D-Lib Magazine. 21, 11/12 (2015), https://doi.org/10.1045/november2015-tkaczyk.
27. Gerasimov, A.N., Elizarov, A.M., Lipachev, E.K.: Subsystem of Formation Metadata for Science Index Databases on Management Platform Electronic Scientific Journals. Russian Digital Libraries Journal. 18(1-2), 6–31 (2015).
28. Gafurova, P.O., Elizarov, A.M., Lipachev, E.K., Khammatova, D.M.: Metadata Normalization Methods in the Digital Mathematical Library. CEUR Workshop Proceedings. 2543, 136–148 (2020).