Metadata Extraction Methods for Organizing a Retro-Collection in the Lobachevskii Digital Mathematical Library

Alexander Elizarov[1, 2][0000-0003-2546-6897], Polina Gafurova[2][0000-0002-1544-155X], Evgeny Lipachev[1, 2][0000-0001-7789-2332]

[1] N. I. Lobachevskii Institute of Mathematics and Mechanics,
[2] Institute of Information Technologies and Intelligent Systems,
Kazan (Volga Region) Federal University
amelizarov@gmail.com, pogafurova@gmail.com,
elipachev@gmail.com

Abstract. Methods for the metadata formation of mathematical retro-digitaized collections and the inclusion of these collections in the digital mathematical library Lobachevskii-DML are proposed. The solutions of a number of metadata management problems arising in the construction of digital mathematical libraries are presented. The services of the metadata factory of the digital mathematical library Lobachevskii-DML, which are used in the formation of digital retro collections, have been developed. The formed archival mathematical collections are described, the features of the metadata of their documents are indicated.

Keywords: Lobachevskii-DML digital math library, metadata factory, metadata management services, archive collections.

1 Introduction

"Like most areas of scholarship, mathematics is a cumulative discipline". One of the program documents of the World Digital Mathematics Library (WDML) Project [1] begins with this phrase. Cumulativeness in this context means that new research is always based on well-organized and well-curated literature. Moreover, the named document provides a comparison of mathematics with art. This comparison is based on the fact that the primary data that mathematicians encounter in their research are human creations, and not data obtained by observation or experiment. All mathematical information that exists today is actually extracted from the mathematical literature or calculated. It is also known that in modern mathematical research, the number of references to documents published in the "pre-digital" period does not decrease. In the works reflecting the activities of many working groups on the integration of mathematical knowledge that are currently functioning, it is noted that both the results obtained earlier

and the systems of reasoning and evidence associated with them should be preserved and made available with the help of modern means of scientific communication [1–4].

Digitization and provision of access to scientific heritage are implemented in various projects.

The JSTOR digital library (https://www.jstor.org/) contains digitized collections not only in mathematics, but also in various fields of knowledge. The volume of this resource is more than 12 million articles of academic journals and books in 75 disciplines. The oldest documents in physics and mathematics are documents from the 15th to 16th centuries.

Projects “Center de diffusion de revues académiques mathématiques” (CEDRAM, http://www.cedram.org/), Gallica (https://gallica.bnf.fr) and Numerisation de Documents Anciens Mathematiques (NUMDAM, http: // www.numdam.org/) form the basis of The French Ecosystem of access to mathematical documents in French, both archival and created in digital format [5, 6].

The project “All-Russian Portal Math-Net.Ru” has been developing since 2006 [7–9]. Currently, the portal of this project presents the digitized archives of the leading Russian mathematical journals, starting from the moment of their opening. For example, the issues of the journal “Matematicheskii Sbornik” have been presented since 1866.

Scientific digital libraries as a specialized class of information systems are the most important component of any scientific information space, and the construction of such libraries is directly aimed at integrating knowledge and expanding access to them (for example, [10]). The above is fully reflected in the main tasks set in projects related to the development of digital mathematical libraries and performing functions of integrating mathematical knowledge (for example, [11–14]).

Many of the existing digital mathematical libraries are built as national ones. For this reason, they have peculiarities both in architecture and in the technologies for managing scientific content used in them. An overview of the specifics and functionality of a number of existing digital mathematical libraries is contained, for example, in [14].

Since 2017, the Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/; see also [15, 16]) has been created at Kazan University in accordance with the basic principles of WDML. One of the basic directions of research developed within the framework of this digital mathematical library is associated with the development of a system of interconnected software tools that provide: creation, processing, storage, management of metadata of digital library objects and the integration of created digital collections into digital scientific libraries that aggregate them. The system of such tools is called by us a metadata factory (see [17]). A description of this system is provided in the next section.

Kazan University Nikolai Lobachevskii Scientific Library including 15,000 manuscripts and 3,000 rare books. It contains unique archives of mathematical documents of the 19th and 20th centuries. Some of the archives are in the process of being digitized. However, these archives are not designed as digital collections. Moreover, no metadescription of the documents that make up digital archives has been created. Difficulties in the formation of metadata sets of digital retro-archives are associated with the need

to develop special methods for their formation and to customize the metadata factory tools for managing them.

In this article, we present those services of the Lobachevskii-DML that we can apply to the mathematical retro-collections. We present the created mathematical retro-digitized collections. We also note the specificity of retro collections of mathematical documents.

2 Mathematical Documents Management Techniques in the Digital Mathematical Library

The creation of a digital mathematical collection and digital library involves solving a number of time-consuming tasks. Content management software tools are an essential component of any digital library. Many of these tools are used by the metadata factory to create, process, store, and manage metadata for digital documents. These tools allow integrating the created digital collections into aggregating digital scientific libraries. Let's describe in more details the available solutions.

Digital libraries, as well as scientific knowledge aggregators, offer a range of software tools for working with content. First of all, these are services for searching in digital collections. For example, a semantic document search facility is available on the European Digital Mathematical Library (EuDML) project site (https://initiative.eudml.org/) [18].

Some steps to optimize the tools of the metadata factory:

- identifying the features of the metadata of documents from various digital collections. These features are caused by the formats of documents, the completeness of the set of metadata, the refinement of the set during the life cycle of a scientific publication;
- select scientific content management tools;
- adapt methods for integrating repositories with other information systems and aggregating libraries.

As a result, we developed the following tools for the metadata factory of the Lobachevskii digital mathematical library (see also [17]):

- tools for the automated generation of metadata for digital mathematical collections;
- xml notation for metadata representation retro-digitized mathematical collections. This notation based on the Journal Archiving and Interchange Tag Suite (NISO JATS) [19];
- software for normalization the metadata of electronic collections of scientific documents according to xml schemes of aggregating libraries in Mathematics and Computer Science;
- converting metadata algorithm for import into the DSpace digital storage;
- integration methods of existing digital mathematical collections of Kazan University into digital mathematical libraries [20, 21].

As in the case of any digital scientific library, the organizing of the Lobachevskii-DML library and the corresponding metadata factory required the involvement of previously created ones, as well as the development of new technological solutions for content management.

At the stage of preprocessing, we did not consider those documents, the processing of which is not supported by the tools of the metadata factory in automatic mode. Together with the document itself, we loaded the reference information about the document into the metadata factory, for example, about its type and encoding. The main objects that the metadata factory processes are article files in various formats. Therefore, one of the goals of preprocessing is to determine the type of document: article, monograph, or collection of articles. Further steps are performed for articles and monographs. Collections of articles are divided into separate articles programmatically (based on the structural features of the document). These articles are then sent for processing to the metadata factory. One of the approaches to solving this problem is described in [22].

Note that the reasons why some documents might not be processed in the metadata factory are reflected in the report that is automatically generated.

At the stage of metadata extraction, full texts of documents are processed and templates for searching for required metadata are used. Also, at this stage, some spelling errors are corrected that occur when extracting metadata from texts obtained as a result of recognizing digitized documents. It also corrects the erroneous case selection and removes unnecessary spaces and characters. Note that the extraction stage is one of the basic stages of the metadata factory functioning.

Metadata Extraction Services are responsible for extracting metadata from documents and external resources. The extraction of basic metadata at the first stage of extraction essentially depends on their explicit presence in the document. Text analytics tools are also used to extract information from text. Collections of digital documents, which include the article under consideration, as well as Internet resources can be used as external resources.

The wide distribution of metadata of various documents on the Internet has led to the fact that one of their sources can be the web pages of the metadata aggregator site or the digital library itself. Thus, when forming a set of metadata for documents in digital collections, as well as when obtaining additional metadata, it is necessary to use the metadata stored on external resources. This task is associated with the tasks of finding information in aggregated databases and digital libraries, some of which are partially closed for access or interrupt the connection, allowing only a limited amount of metadata to be downloaded. When searching for metadata on the pages of aggregator sites, one should also take into account that the choice and search order in such sources should be determined in advance, since some sources store information only on a specific topic (for example, the DBLP bibliographic database - on computer topics) or incomplete list of metadata. The peculiarity of this stage is that access mode is also limited to some sites. However, many resources provide the ability to legally extract metadata using the API and the OAI-PMH server. The main steps of the algorithm for extracting metadata from an Internet resource using the example of one of the collections are given in [21].

At the verification stage, the completeness and compliance of the selected metadata with the established rules, written in the form of DTD files or XML schemas, is checked. After passing this stage, there are three options for further actions:

- repeated extraction of necessary and additional metadata, as well as their verification;
- issuing a report that the metadata factory tools are not sufficient to obtain the required set of metadata;
- transition to the final stage – metadata normalization.

Extraction of additional metadata is aimed at extracting metadata from sources located outside the processed document. These sources include collections, which include the processed document, as well as Internet resources.

A number of special tools of the digital math library metadata factory are designed to perform metadata harmonization and normalization procedures.

Harmonization of metadata implies the possibility of the simultaneous use of several different metadata standards in one software system. Metadata normalization methods are used to map several different metadata standards into a single schema or structure for further use in a single software system (see, for example, [21, 23, 24]).

Tasks related to the normalization of metadata in various formats are one of the most relevant when working with a metadata factory. Examples of such tasks are:

- normalization to formats for internal storage and loading into a digital library;
- normalization to formats of other digital libraries and aggregators or presentation in the form of bibliographic citation formats.

The Lobachevskii-DML digital library implements several services that normalize metadata to various formats. So, for example, one of them is the service for converting metadata of the digital collection of articles of the Russian Digital Libraries Journal (https://elbib.kpfu.ru/) into the DBLP database format (https://dblp.uni-trier.de/). The developed metadata transformation algorithm includes semantic transliteration of the names and surnames of the authors of the articles. The initial sets of metadata used in the conversion to the named format were generated automatically with the help of software tools developed by us, taking into account the specifics of the Open Journal Systems software platform [25] on which this journal operates. The algorithm for translating this metadata into the DBLP format has been successfully implemented; it is described in detail in [17, 21].

3 Methods of Arranging Mathematical Retro-Collections

As the first example of the formation of archival collections, we will describe the results of the creation of the digital collection “Proceedings of the N.I. Lobachevskii Mathematical Center”, obtained using the services of the metadata factory. This collection has been published since 1998, its main purpose – the publication of materials of mathematical conferences. As a result, most of the volumes of “Proceedings ...” contain several dozen articles with a limited (from a modern point of view) metadata content.

At the same time, during the time that has passed since the release of the first volume, several style rules for the preparation of materials were used, which affected the design of articles and file formats of the collected collections. The necessary conditions for creating a digital collection from the “Proceedings ...” file array were the division of volumes into separate articles, the extraction of metadata describing each article, the generation of additional metadata (containing, in particular, the bibliographic description of the article, a link to the article file in the digital collection, as well as links with the profiles of the authors of the article on academic portals and scientometric databases (kpfu.ru, MathNet.ru, Scopus, etc.). The developed algorithm is presented in [22].

At the next stage, metadata was generated in the format of loading the digital storage DSpace. The file upload format in DSpace is based on the Dublin Core standard. The files are formed in a specific order in such a way that the structure required for loading into DSpace is obtained. The formed service was tested on the collection of “Proceedings ...”, and the corresponding digital collection is included in the digital library Lobachevskii-DML.

Another mathematical archive of undoubted scientific interest is the Archive of the Physics and Mathematics Society of Kazan University. It is based on the issues of the journal “Proceedings of the Physics and Mathematics Society at Kazan University” for 1891–1949. This edition published the leading mathematicians of Russia, and later – of the Soviet Union. Among the authors of the journal articles are outstanding mathematicians M.G. Krein, A.A. Markov, N.G. Chebotarev and N.G. Chetaev.

Since before the formation of this digital collection, the archive was stored only on paper, it was necessary not only to carry out procedures for extracting the metadata of the collection documents, but also to digitize the journals themselves.

Let's highlight the main features of the created “Archive ...”.

Depending on the year of publication, the collections of the archive materials have different styles of articles. At the same time, they practically lack the information necessary to form a fundamental set of metadata according to the EuDML scheme [26]. The difficulties in extracting metadata from articles are illustrated in Figure 1. Here is one of the articles in the archive, which has only a title, and the author is indicated only on the last page. This is the famous article by A.A. Markov “Extension of the law of large numbers to quantities that depend on each other”, published in issue 4, 1906, “Proceedings of the Physics and Mathematics Society at Kazan University”. This article investigates sequences of random events, which are now commonly called Markov chains.

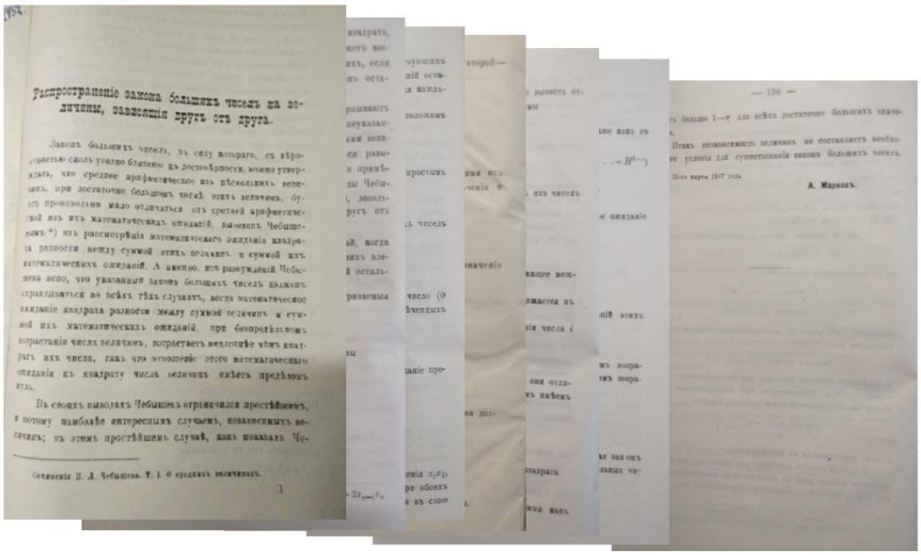

Fig. 1. Lack of information in the articles necessary to form a set of metadata, in particular, information about the authors, keywords.

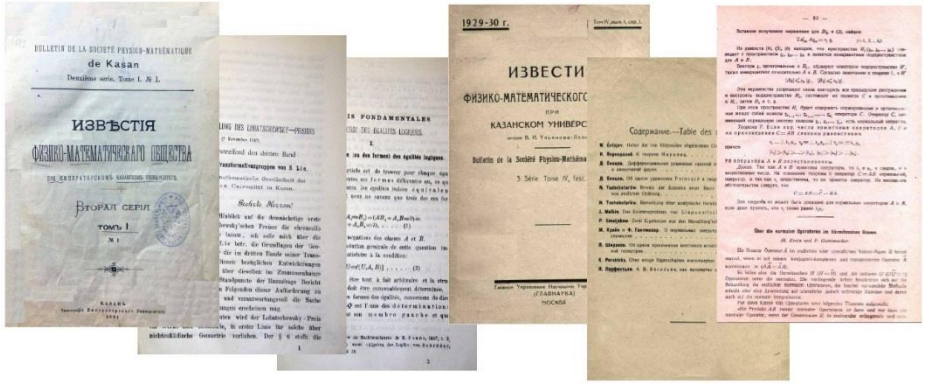

Fig. 2. Articles in one issue can be presented in Russian, French, German, English or Italian. Articles in Russian are accompanied by a summary in one of the listed European languages.

One collection may contain articles in several languages; at the same time, for articles in Russian, a mandatory requirement of the editors was the inclusion of a summary in one of the following languages: English, German, French or Italian (see Fig. 2).

Let's highlight the completed stages of work with the named archive collection.

– **Stage 1.** Creation of a meta-description of the archive of journal articles in formats that allow machine processing. It was assumed that the meta description should include a bibliographic record of all articles of the specified journal. Since the journal

was not digitized, this stage could not be automated. An additional difficulty that has arisen is the need to work with the library paper fund only using the system of extracts from the catalog.

- **Stage 2.** Presentation of the digital collection in the digital library Lobachevskii-DML in the form of a system of metadata and links to the catalog of the Scientific Library of Kazan University.
- **Stage 3.** Organization of digitization of the archive of the specified journal.
- **Stage 4.** Formation of a digital collection, including full texts of articles of the specified journal, provided with metadata sets in the Lobachevskii-DML, MathNet.ru formats and the format of the mandatory metadata set of the European Digital Mathematical Library (EuDML, https: // initiative. eudml.org/).
- **Stage 5.** Inclusion of the generated digital collection in Lobachevskii-DML with a set of metadata and full texts of articles.

The machine processing of retro-archive articles and the formation of metadata were carried out in accordance with the following algorithm.

Algorithm 1: Extraction and normalization metadata

1: read pdf article files
2: highlight the title of the article
3: determine the main language of the article
4: identify the author of the article
5: translate the title and author into modern Russian (in case the original language is pre-reform Russian)
6: clarify information about the author from open Internet sources
7: extract the annotation
8: define the language of the annotation
9: extract information about author and title in the annotation language
10: extract the first and last page numbers of an article
11: extract bibliography
12: refine and supplement article information
13: forming a set of metadata about the article

4 Conclusion

The article describes the structure of the metadata factory of the digital mathematical library Lobachevskii-DML and the set of services that form the basis of this factory. The basic services of the Lobachevskii-DML digital library metadata factory are presented that can be applied to archival mathematical collections. The description of the formed archival mathematical collections is given, the features of the metadata of their documents are indicated.

The further direction of development is seen in the improvement of the created metadata factory and the development of possibilities for its use in any scientific digital libraries.

The work was carried out within the framework of the development program of the Regional Scientific and Educational Mathematical Center of the Volga Federal Region, agreement number 075-02-2020-1478/1.

References

1. Developing a 21st Century Global Library for Mathematics Research. The National Academies Press, Washington (2014).
2. Ion, P.: The Effort to Realize a Global Digital Mathematics Library. In: Greuel G.M., Koch T., Paule P., Sommese A. (eds). Mathematical Software – ICMS 2016. ICMS 2016. Lecture Notes in Computer Science, vol. 9725. Springer, Cham (2016), https://doi.org/10.1007/978-3-319-42432-3_59
3. Ion, P.D.F., Watt, S.M.: The Global Digital Mathematics Library and the International Mathematical Knowledge Trust. In: ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence, vol. 10383, pp. 56–69. Springer (2017), https://doi.org/ 10.1007/978-3-319-62075-6_5.
4. Rocha, E.M., Rodrigues, J.F.: Disseminating and preserving mathematical knowledge. In: Borwein, J.M., Rocha, E.M., Rodrigues, J.F. (eds.). Communicating Mathematics in the Digital Era, pp. 3–21. A K Peters, Ltd. (2008).
5. Bouche, T.: Toward a Digital Mathematics Library? A French Pedestrian Overview. In: Borwein, J.M., Rocha, E.M., Rodrigues, J.F. (eds.). Communicating Mathematics in the Digital Era, pp. 47–73. A K Peters, Ltd. (2008).
6. Bouche, T., Labbe, O.: The New Numdam Platform. In: CICM 2017: Intelligent Computer Mathematics, pp. 70–82 (2017), https://doi.org/10.1007/978-3-319-62075-6_6.
7. Zhizhchenko, A.B., Izaak, A.D.: The information system Math-Net.Ru. Application of contemporary technologies in the scientific work of mathematicians. Russian Math. Surveys, 62(50), 943–966 (2007), http://dx.doi.org/10.1070/RM2007v062n05ABEH004455.
8. Zhizhchenko, A.B., Izaak, A.D.: The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals. Russian Math. Surveys, 64(4), 775–784 (2009), http://dx.doi.org/10.1070/RM2009v064n04ABEH004638.
9. Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., Zhizhchenko, A.B.: Math-Net.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. Xie, I., Matusiak, K.K.: Discover Digital Libraries: Theory and Practice. Elsevier Inc. (2016).
11. Bouche, T.: Some Thoughts on the Near-Future Digital Mathematics Library. Towards a Digital Mathematics Library. Masaryk University, pp. 3–15 (2008), https://eudml.org/doc/221606, last accessed 2020/12/12.
12. Bouche, T.: Digital Mathematics Libraries: The Good, the Bad, the Ugly. Math. Comput. Sci., 3, 227–241 (2010), https://doi.org/10.1007/s11786-010-0029-2.
13. Bouche, T.: The Digital Mathematics Library as of 2014. Notices Amer. Math. Soc., 61(9), 1085–1088 (2014).
14. 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).

15. Elizarov, A.M., Lipachev, E.K.: Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University. CEUR Workshop Proceedings, 2022, 326–333 (2017).
16. Elizarov, A., Lipachev, E.: Big Math Methods in Lobachevskii-DML Digital Library. CEUR Workshop Proceedings, 2523, 59–72 (2019).
17. 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), https://doi.org/10.26907/1562-5419-2020-23-3-336-381.
18. D7.4: Toolset for Image and Text Processing and Metadata Enhancements – Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/images/D7.4-v1.0.pdf, last accessed 2020/12/12.
19. Journal Article Tag Suite. https://jats.nlm.nih.gov/about.html, last accessed 2020/12/12.
20. Elizarov, A.M., Zaitseva, N.V., Zuev, D.S., Lipachev, E.K., Khaidarov, S.M.: Services for Formation of Digital Documents Metadata in the Formats of International Science-based Databases. CEUR Workshop Proceedings, 2260, 175–185 (2018).
21. 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).
22. Elizarov, A.M., Lipachev, E.K.: Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library. CEUR Workshop Proceedings, 2543, 354–360 (2020).
23. Nilsson, M., Naeve, A., Duval, E., Johnston, P., Massart, D.: Harmonization Methodology for Metadata Models. https://hal.archives-ouvertes.fr/hal-00591548, last accessed 2020/12/12.
24. Nilsson, M.: From Interoperability to Harmonization in Metadata Standardization. Doctoral thesis, Stockholm, Sweden (2010).
25. MacGregor, J., Stranack, K., Willinsky, J.: The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication. In: Bartling, S., Friesike, S. (eds). Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, pp. 165–175 (2014), https://doi.org/10.1007/978-3-319-00026-8_3.
26. EuDML metadata schema specification (v2.0–final), https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2020/12/12.