Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library

A. M. Elizarov[0000-0003-2546-6897] and E. K. Lipachev[0000-0001-7789-2332]

N. I. Lobachevskii Institute of Mathematics and Mechanics,
Higher School of Information Technologies and Intelligent Systems,
Kazan (Volga Region) Federal University
amelizarov@gmail.com, elipachev@gmail.com

Abstract. Digital mathematical libraries are today one of the tools for integrating mathematical knowledge. This integration method is based on the use of metadata. Our task was to create methods to programmatically extract the necessary objects from digital mathematical documents, establish semantic relationships between them and generate the necessary sets of metadata. Based on the analysis of the structure of the set of documents under consideration and the stylistic features of their design, an algorithm has been developed for extracting their metadata, creating digital collections and then including them in the corresponding digital library. The algorithm is implemented as a software system and tested on the example of a set of files "Proceedings of the N.I. Lobachevskii Mathematical Center" for 1998–2018. A corresponding digital collection has been created, which is included in the Lobachevskii Digital Mathematical Library (Lobachevskii DML, https://lobachevskii-dml.ru/).

Keywords: Digital Collection, Digital Mathematics Library, Metadata, Semantic Relation, Semantic Method, Lobachevskii DML.

1 Introduction

Currently, one of the tools for integrating mathematical knowledge is digital mathematical libraries (see, for example, [1]–[5]). This integration method, as well as information management on the Web, is based on the use of metadata [6]–[8]. Creating a new digital collection from a set of files containing, for example, an archive of journal articles, involves a series of operations to coordinate formats, extract and refine metadata, and normalize them in accordance with established data schemes. Each such operation requires the use of special methods and software tools that take into account the specifics of the collection being processed and the rules for organizing a digital library. When creating digital libraries, additional requirements are also imposed, in particular, on the composition and format of metadata [9]. The construction of digital mathematical libraries requires the development of software tools that take into account such features of mathematical documents as the presence of formulas, notation, definitions, theorems, and proofs in texts. All of the above form a complex

system of objects, interconnected both within the framework of the document under consideration, and with other documents and objects in this field of mathematics [1, 4]. Our task was to create methods that allow the software to extract the necessary objects from digital mathematical documents and establish semantic relations between them [10–14].

Periodic journals with a long history of publication have archives of articles that differ significantly in the composition of the metadata presented in these articles. In addition, style rules, fixed, for example, by MS Word templates or in `.sty`-files (for the TEX system), have also changed many times over the past. For example, in the journal "Lobachevskii Journal of Mathematics" (the articles of this journal constitute one of the collections of the digital mathematical library Lobachevskii DML) from 1998 to 2019, foor different `.sty`-files were used to design articles. Algorithms for extracting metadata use such stylistic features as heading, annotation fonts, and the document structure that determines the sequence of structural blocks (name, authors list, etc.) (see, for example, [15–17]). Algorithms for extracting metadata use such stylistic features as fonts for the title of the article, its annotation and a set of keywords, as well as the structure of the document, which determines the sequence of structural blocks (name and surname of authors, list of authors, etc.) (see, for example, [15–17]).

Automated processing of arrays of scientific documents is complicated by the variety of styles used in journals. As a result, various methods of extracting metadata are required, which will take into account the structural features of a particular collection (see, for example, [10, 18]).

This work is devoted to the development of methods for creating digital scientific collections from an array of heterogeneous digitized documents. On the example of processing a set of files containing volumes of "Proceedings of the N.I. Lobachevskii Mathematical Center" (hereinafter referred to as the "Proceedings") for 1998–2018, describes the process of forming the corresponding digital collection and its inclusion in the digital mathematical library Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/), which is currently being formed in Kazan (Volga region) Federal University.

The main purpose of this "Proceedings" is the publication of materials of mathematical conferences. As a result, most volumes of the "Proceedings" contain dozens of articles with a limited (from the modern point of view) metadata composition. Since 1998 (the moment the first volume was released), several style rules for the preparation of materials have been used, which affected the design of articles and the choice of file formats of compiled collections.

Let us single out the main tasks of forming a digital collection. The necessary conditions for creating a digital collection from the "Proceedings" array were:

- division of volumes into separate articles;
- highlighting metadata describing each article;
- generation of additional metadata that contains, in particular, a bibliographic description of the article, a link to the article file in the digital collection, as well as links to the profiles of article authors on academic portals and scientometric databases (kpfu.ru, MathNet.ru, Scopus, DBLP and etc.).

2 Algorithm for Extracting Metadata and Semantic Relationships

Let us describe in more detail the main stages of the software processing of the selected set of files of collections of articles that have been processed. These stages are as follows:

- highlighting metadata;
- their conversion in accordance with XML schemas;
- creation of a digital collection;
- its inclusion in the digital mathematical library Lobachevskii-DML.

2.1 Clustering

First of all, file clustering was performed. As a result, the corresponding volumes were divided into classes according to the similarity of their structure and design. From the moment the first volume was released (1998) until the publication of the last 57th volume (to date), various rules were used to prepare the materials. This affected the file formats of the volumes, as well as the design of the articles themselves. For example, only a number of documents contain indexes of the Universal Decimal Classification (UDC). Table 1 shows the differences in the use of structural elements and the composition of the metadata of this collection.

Table 1. Metadata and building blocks in collection volumes.

e-mail	UDC	Contents	Author Index
3, 4, 12, 15, 18,19, 22, 24, 31, 35-40, 42–47, 49, 50, 52–57	54, 55, 56, 57	4, 9, 15, 17, 20, 22, 26, 29, 32, 33, 49	2, 3, 5, 6, 8, 9, 12, 13, 16, 18, 19, 21, 23, 24, 28, 31, 34–40, 42–47, 50, 52–57

2.2 Metadata Extraction

Further, in order to extract metadata describing both the volume as a whole and the articles included in it, the collection files were processed. In particular, the page numbers of all the articles of each volume were determined. To search for pages with article titles, an algorithm has been developed that uses the structural homogeneity of each volume and style uniqueness in the design of articles in it.

An important part of the structural analysis of documents is the allocation of blocks such as the name, surnames of the authors, their affiliation, abstract, key words and bibliographic records. Special ontologies have been developed to describe the structure of scientific documents [19, 20]. For the semantic structuring of digital content, they use ontologies CiTO, DoCo, SWAN, SKOS, CERIF and SPAR (see, for example, [21, 22]). An example of the distribution of such blocks according to structural features and their description in terms of DoCO ontologies is given in [10].

To extract article metadata based on characteristic features, we have defined rules for selecting article blocks. Such features include, in particular, the design of article styles (font, size, use of selections, and a number of others). Improving the quality of metadata extraction provides some additional features that take into account:

• text structuring (for example, the location of the word "Annotation" in front of the annotation block);

• the type of email address record template used;

• the position of the block in the text (for example, the document begins with the title of the article).

As the main elements of the article taken into account by these functions, you can use the position of the block in question in the document, as well as the font used in the text of this block. These functions made it possible to distinguish not only the titles of articles, but also lists of authors, bibliography blocks, and other metadata (for example, e-mail, keywords) if they exist in the text.

Using text analysis methods [23, 24], the terms from which the sets of keywords were formed for inclusion in the metadata were formed from the documents of the digital collection.

A number of metadata (such as email addresses of authors, their affiliation) was imported and updated from the profiles of authors on academic sites and in scientometric databases. In this procedure, the semantic relationships established during the formation of the digital collection were applied.

2.3 XML-representation

We have developed an XML language for describing digital mathematical collections, consisting of a set of tags and XML schemes based on the Journal Archiving and Interchange Tag Suite (NISO JATS, https://jats.nlm.nih.gov/archiving/). In the notation of this language, based on the data obtained during the processing of an array of files, a description of the collection of "Proceedings" is performed.

2.4 Splitting Volume Files into Article Files and Creating a Digital Collection

The next stage in creating the digital collection consisted of the procedures for dividing each volume of the "Proceedings" into separate articles. To do this, tags whose attributes indicate the start and end pages of the articles were read from XML-files that contain meta descriptions of volumes. After that, the files were divided into separate documents, which were named in accordance with the rules of the digital collection.

The system of metadata prepared in the process of the above algorithm allowed to form a digital collection of "Proceedings of the N.I. Lobachevskii Mathematical Center" and include it in the digital library Lobachevskii DML (https://lobachevskii-dml.ru/) [11].

2.5 Software Implementation

The algorithm is implemented in the form of programs in C #, allowing you to process files in the formats TeX, OpenXML (`.docx`) and `.pdf`. TeX-files were generated using standard functions that implement operations with text strings. To work with pdf-files, we used the functions of the PDFLib libraries (https://www.pdflib.com) and iTextSharp (https://www.nuget.org/packages/iTextSharp/). For documents presented as docx-files, the "`word/document.xml`" file was parsed from the .docx archive in accordance with the Office OpenXML format (see, for example, [25]).

The process of selecting articles was carried out using a program developed in Python using the functions of the PyPDF2 library (http://pybrary.net/ pyPdf/).

3 Conclusion

For inclusion in the international scientific space of digital mathematical collections of Kazan University, methods of their formation from a set of documents presented in various storage formats are proposed. Based on the analysis of the structure of documents and the stylistic features of their design, an algorithm for the extraction of their metadata has been developed, implemented on the example of "Proceedings of the N.I. Lobachevskii Mathematical Center".

The work partially contains the results 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 of the National Technological Initiative "Center for storing and analyzing big data", supported by the Ministry of Science and Higher Education of the Russian Federation under the Treaty of Moscow State University named after M.V. Lomonosov with the Project Support Fund of the National Technological Initiative dated 15/08/2019 No. 7/1251/2019. The work was also carried out with the partial support of the Russian Fund for Basic Researches (project 18-29-03086); with the partial support of the Russian Fund for Basic Researches and the Government of the Republic of Tatarstan within the framework of scientific project 18-47-160012.

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: G.-M. Greuel et al. (Eds.). ICMS 2016, LNCS 9725. Springer International Publishing Switzerland, 458–466 (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. Elizarov, A.M., Lipachev, E.K., Zuev, D.S.: Digital Mathematical Libraries: Overview of Implementations and Content Management Services. CEUR Workshop Proceedings, vol. 2022, pp. 317–325 (2017).
5. Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., and 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.
6. Gartner, R.: Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer (2016).
7. Sicilia, M.-A. (Ed.): Handbook of Metadata, Semantics and Ontologies. World Scientific Publishing Co. Pte. Ltd. (2014).
8. Lubas, R., Jackson, A., Schneider, I.: The Metadata Manual. Chandos Publishing (2013).
9. Alemu, G., Stevens, B.: An Emergent Theory of Digital Library Metadata. Elsevier Ltd. (2015).
10. Elizarov, A.M., Khaydarov, Sh.M., Lipachev, E.K.: Scientific Documents Ontologies for Semantic Representation of Digital Libraries. 2nd RUSSIA AND PACIFIC CONF. ON COMPUTER TECHNOLOGY AND APPLICATIONS, pp. 1–5 (2017), https://doi.org/10.1109/RPC.2017.8168064.
11. Elizarov, A.M., Lipachev, E.K.: Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University. CEUR Workshop Proceedings, vol. 2022, pp. 326–333 (2017).
12. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A., Solovyev, V.D., and Zhiltsov, N.G.: Mathematical knowledge representation: semantic models and formalisms. Lobachevskii J. of Mathematics, 35 (4), 348–354 (2014), https://doi.org/10.1134/S1995080214040143.
13. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A.: Mathematical Knowledge Management: Ontological Models and Digital Technology. In: CEUR Workshop Proceedings, vol. 1752, pp. 44–50 (2016).
14. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A.: Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management. In: Communications in Computer and Information Science, vol. 70, pp. 33–46. Springer (2017), https://doi.org/10.1007/978-3-319-57135-5_3.
15. Chen, J., Chen, H.: A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning. Journal of Software, vol. 8, no. 1, pp. 55–62 (2013), https://doi.org/10.4304/jsw.8.1.55-62.
16. Ronzano, F., Saggion, H.: Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. In: Japkowicz N., Matwin S. (eds) Discovery Science. Lecture Notes in Computer Science, vol 9356, Springer, Cham. (2015), https://doi.org/10.1007/978-3-319-24282-8_18.
17. Tkaczyk, D., Tarnawski, B. and Bolikowski, Ł.: Structured Affiliations Extraction from Scientific Literature. D-Lib Magazine, vol. 21, no. 11/12 (2015), https://doi.org/10.1045/november2015-tkaczyk.
18. Elizarov, A.M., Lipachev, E.K., and Khaydarov, S.M.: Automated system of services for processing of large collections of scientific documents. CEUR Workshop Proceedings, vol. 1752, pp. 58–64 (2016).
19. Peroni, S.: Semantic Web Technologies and Legal Scholarly Publishing. Springer International Publishing, (2014), https://doi.org/10.1007/978-3-319-04777-5.

20. Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F.: The Document Components Ontology (DoCO). Semantic Web, vol. 7, no. 2, pp. 167–181 (2016), https://doi.org/10.3233/SW-150177.
21. Ruiz-Iniesta, A., and Corcho, O.: A review of ontologies for describing scholarly and scientific documents. CEUR Workshop Proceedings, vol. 1155, pp. 1–12 (2014).
22. Kogalovsky, M.R., Parinov, S.I.: Scholarly Communication in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships. In: Klinov P., Mouromtsev D. (eds) Knowledge Engineering and Semantic Web. Communications in Computer and Information Science, Springer, vol 518, pp. 87–101 (2015), https://doi.org/10.1007/978-3-319-24543-0_7.
23. Ingersoll, G. S., Morton T. S., Farris A. L.: Taming Text. How to Find, Organize, and Manipulate It. Manning Publications Co. (2013).
24. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC. Education Services (Ed), Wiley (2015).
25. Standard ECMA-376 Office Open XML File Formats, http://www.ecmainternational.org/publications/standards/Ecma-376.htm. last accessed 2019/11/21.