Scientific Documents Ontologies for Semantic Representation of Digital Libraries

Alexander Elizarov, Shamil Khaydarov, Evgeny Lipachev
N.I. Lobachevskii Institute of Mathematics and Mechanics
Kazan (Volga Region) Federal University
Kazan, Russian Federation
amelizarov@gmail.com, 15jkeee@gmail.com, elipachev@gmail.com

***Abstract*— We present a system of services for the automatic processing of collections of scientific documents that are part of digital libraries. These services are based on ontologies for scientific documents representation, as well as on methods for semantic analysis of mathematical documents. The developed tools automatically check validity of documents for compliance with manuscript guidelines, convert these documents into required formats and generate their metadata.**

Keywords—automated processing of scholarly papers; document semantics; metadata extraction; ontology; semantic publishing

I. INTRODUCTION

The transition to the digital form of information presentation requires not only the application of new formats for the representation of the document itself, but also additional data about it (including metadata) with a view to further processing them throughout the life cycle of the Scientific Publication [1, 2]. The process of forming digital documents differs from the traditional processes of preparing paper publications. These differences are so great that the transition to a digital form of information is usually called the "digital revolution". Traditionally, digital scientific documents are stored in unstructured form. For this reason, these documents are difficult to process and classify [3-5].

One of the modern requirements for digital documents is the possibility of their semantic processing. For such processing special markup languages are used, as well as ontologies for formalizing existing links. In order for a scientific document to be machine-readable, each of its elements must be marked with special labels that form the so-called document metadata. Currently, there are many options for storing documents with metadata. The most popular of these are the languages of semantic markup. One way to formalize such markup is to use ontologies. There are several ontologies designed for document structure representation [6, 7].

In this paper, we present a new method for the formation of semantic representation of documents included in the digital scientific collection. This method is based on an analysis of the structure of documents and their stylistic features. Thus, an attempt has been made to relate two technologies: the processing of unstructured data and their convertation into a machine-readable form. Section 2 presents methods for processing scientific documents based on an analysis of their structure. The result of applying these methods is semantic representation of these documents. In Section 3 the ontologies used to describe the structure of documents are characterized. Section 4 presents methods for formation of a semantic representation of scientific documents. As an example of the application of these methods, the result of automatic processing of a large collection of works (containing more than 1500 documents) of the XI All-Russian Congress on Fundamental Problems of Theoretical and Applied Mechanics (Russia, Kazan, 2015) is presented. Section 5 presents the services intended for the preparation of document's final layout version for print. Section 6 describes the method of automated metadata formation of scientific publications for the Russian Science Citation Index.

This work was financed by a grant allocated to the Kazan Federal University for the state task in the field of scientific activity (grant agreement No. 1.2368.2017), as well as Russian Foundation for Basic Research and Government of the Republic of Tatarstan in the framework of scientific projects No. 15-07-08522, 15-47- 02472.

II. PROCESSING OF UNSTRUCTURED DATA

Traditionally, scientific publications can be presented in layout-oriented formats (such as .pdf, .docx, .tex). Managing collections of such documents requires the development of additional tools and services. The most typical tasks are the following: organization of search services, document clustering, search for related articles, and others. To implement these services, you need to use text analytics tools, in particular, methods for separating metadata based on the structure of documents [3–5, 8–13]. Typically, the technique of regular expressions is used to extract metadata. For example, in [3, 12] it was suggested to create metadata of scientific publications using regular expressions, as well as information about formatting (what font is used, what is its size, whether it is an underscore, etc.). The corresponding algorithm is reduced to converting documents from .docx, .pdf formats to documents with xml-markup and syntactic analysis of text based on the results of its formatting.

A number of works on extracting structured data from scientific documents are also well known (see, for example, [3, 10, 11]). Thus, the work [10] describes a method for extracting data from documents in the pdf-format. This method translates a pdf-document into a "semi-semantic" text, and then it performs its annotation. In the same work, a method for storing documents in the pdf-format is proposed, which allows to translate such documents into the rdf-format.

III. ONTOLOGIES FOR DESCRIBING DOCUMENTS STRUCTURE

There are several ontologies for describing scientific documents such as [1, 2, 6, 7, 12, 14]. For the semantic structuring of digital content, the ontologies CiTO, DoCo, SWAN, SKOS, CERIF and SPAR are used in them [7, 14]. The later ontology, which stands for Semantic Publishing and Referencing Ontologies, is designed to describe journal articles and books [1, 2].

This suite of ontologies consists in independent re-usable ontologies designed for describing the semantics of bibliographic objects, as well as their citation references, bibliographic records, components of documents and various aspects of the scientific publication process. In fact, all these ontologies are taxonomies and are described in the OWL2 DL and RDF languages developed by the W3C consortium. The first four of them (FaBiO, CiTO, BiRO and C4O) are for representation of bibliographic objects, bibliographic records and sources in the lists of literature contained in publications, quotations, citation contexts and their references to the relevant sections of cited publications, and for organizing bibliographic records and references in bibliographies, ordered lists of sources and library catalogs. The remaining ontologies (DoCO, PRO, PSO and PWO) serve to create structured managed dictionaries for document components, publishing roles, publishing states, and workflows in publishing processes.

DoCO ontology provides a broad number of classes and relationships that allow describing a document based on its structure and content. For example, it is describe the vast majority of document components such as chapter, preface, glossary, etc. DoCO imports two ontologies: DEO and the Ontology of document structural. DEO is an OWL2-ontology that describes the main rhetorical elements of the document. It also provides a structured vocabulary for rhetorical elements within documents and it uses all the rhetorical block elements from the SALT Rhetorical Ontology. The ontology of templates formally defines templates for segmenting a document into atomic components so that they can be used independently in different contexts [1, 2, 7].

An important part of the structural analysis of documents is the allocation of such blocks as the name, authors' surnames, their affiliation, abstract, keywords and bibliographic records. Table 1 gives an example of the distribution of such blocks by structural features and their description in terms of DoCO ontologies. This table presents a set of features used by us in analyzing the collection of materials of the XI All-Russian Congress on Fundamental Problems of Theoretical and Applied Mechanics (August 20–24, 2015, Kazan, Russia).

TABLE 1. A SET OF FEATURES USED IN THE ANALYSIS OF THE COLLECTION OF MATERIALS OF THE XI ALL-RUSSIAN CONGRESS ON FUNDAMENTAL PROBLEMS OF THEORETICAL AND APPLIED MECHANICS

Paper block	Block feature	Ontology concepts
Title	Font: Times New Roman, 12 pt, bold, centered etc. Position: at the start of the document	doco:title
Author's list	Font: Times New Roman, 12 pt, centered etc. Position: after title Regexp Pattern: authors separated by comma	doco:ListOfAuthors, feof:author
Affiliations	Font: Times New Roman, 12 pt, italic, centered etc. Position: after author's list	pro:relatesToOrganization
E-mail	Font: Times New Roman, 9 pt, bold, centered etc. Position: after affiliations Regexp Pattern: Unique address type	fabio:Email
Abstract	Font: Times New Roman, 9 pt, justified etc. Position: after e-mail Regexp Pattern: Begins with a specific word: abstract	doco: abstract
References	Posttion: at the end of the document Regexp Pattern: Begins with a specific word: References	doco:bibliography, deo:BibliographicReference

For mathematical documents, you can also use ontology of mathematical knowledge OntoMathPRO [15, 16].

IV. METHOD OF FORMATION OF SEMANTIC REPRESENTATION OF SCIENTIFIC DOCUMENTS

Machine-oriented processing of electronic collections assumes the presence of semantic markup of their documents. Such markup can be done in automatic mode on the basis of information about the structure of each document and the features of its formatting.

All the documents of the collection we are working with were first converted to the OpenXML format (see [17]). Further, structural clustering of these documents was carried out. For this purpose the collection was divided into classes of documents similar in structure. Then the semantic representation of each document was formed: for each class of documents, a pattern of regular expressions was chosen with the possibility of adjusting them in the process of work. Tools have also been developed that use these patterns to highlight

information blocks (title of the article, list of authors, block of literature, etc.).

The service system includes modules that perform the following functions:

- extracting metadata from the collection documents based on the analysis of the structure of documents and formats of information presentation;
- automatic selection of documents in accordance with the established order, for example, lexicographical, or according to the lists of authors;
- extraction of annotation blocks from the collection documents, preparation of an alphabetical index and the formation of a collection of annotations;
- automatic generation of a bibliographic description of an article from a collection with the recording of this information in a block of document footers;
- conversion of documents into a pdf-format in accordance with the established parameters;
- formation of the final original layout of the generated edition with automatic selection of articles, placement of pages, preparation of the alphabetical index and content;
- preparation of metadata for export to scientific citation databases.

The process of document validation and the corresponding style casting service involves checking the presence and location of key blocks (title of the article, list of authors, affiliation, keywords, etc.) in accordance with the documents regulating the publishing formats. The developed style cast service implements the following steps:

- a uniform presentation of the title of the articles, lists of authors (for example, S. M. Khaydarov is written instead of S. M. Khaydarov);
- a uniform representation of the authors' affiliation, for example, the entries "KFU", "K(P)FU", "Kazan University", "Kazan (Volga Region) Federal University" and "Kazan Federal University" are brought to a unified form "Kazan (Volga Region) Federal University"; to do this, a dictionary of synonyms is created;
- uniform font design of sections of the text of articles; the register is recorded when recording key blocks, for example, the title of the article is written in capital letters;
- a set of mathematical formulas and a system of references to them;
- bringing the list of literature in line with the chosen bibliographic description format;
- forming links to sources of support for research grants and gratitude.

To extract the metadata of an article based on the characteristic features (see Table 1), we define the rules for selecting the blocks of an article. Such features include, in particular, style design of articles (font, font size, selection, etc.). To increase the quality of metadata extraction, some additional features allow: the patterning of the text (for example, the location of the word "Annotation" before the annotation block or the template type of the e-mail address record) and the position of the block in the text (for example, the document starts with the title of the article). As such signs, the block position in the document can be used, as well as the font used in the text of this block (see, for example, [3, 11, 18–20]). Table 1 shows the set of features that we used in the structural analysis of the collection of scientific documents published in the materials of the 11th All-Russian Congress on Fundamental Problems of Theoretical and Applied Mechanics.

The corresponding program module is implemented as a PHP-script. This module implements the following steps:

- the document.xml file is extracted from the article file in .docx format; further, using the description of the DOMDocument class (see, for example, [21]), this file is parsed;
- the getElementsByTagNameNS method with the parameter "w:p" (the paragraph markup tag in OpenXML) is used to allocate text blocks; as a result, we get the DOMNodeList object, in which all the paragraphs of the document are stored as a list;
- the received list is consistently checked for compliance with the specified rules; This allows you to generate a semantic representation of the document.

Note that in some cases the allocated block can be stored in several paragraphs. This situation is taken into account with the help of an additional analysis of the paragraph that follows the paragraph under consideration.

Regular templates are used to specify a template, for example, to select a list of authors such a template has the form:

«/ ([A-ЯA-Z]\. (?: [A-ЯA-Z]\.)? \S [A-ЯA-Z] [a-za-я]+) (,\s)? (?1)? (,\S)? (?1)? /».

In addition, the presented software module checks the presence of key structures and their correspondence to the specified format. The result of the module is an document containing the metadata of the articles in the collection being processed.

V. CREATION OF DOCUMENT'S FINAL LAYOUT VERSION

We have developed a service for creating the final layout version of the scientific publication document, which allows you to automatically prepare from the files of the electronic collection the original template of the scientific publication (collection of materials, journal and others). The order of articles placement is determined by the semantic view of the collection stored in the XML-file. The algorithm is implemented as a VBA macro and includes the following steps: in the first step, the initial and final page numbers are calculated according to paper position in document's final layout version. Next, we sequentially open the collection

documents in accordance with the order specified in the XML-file in accordance with the extraction rules. We compute the initial and final pages, after which we form the bibliographic description of the article. It is recorded in header and footer of the document. The received document is converted to pdf-format. We also save the bibliographic description in an XML-file.

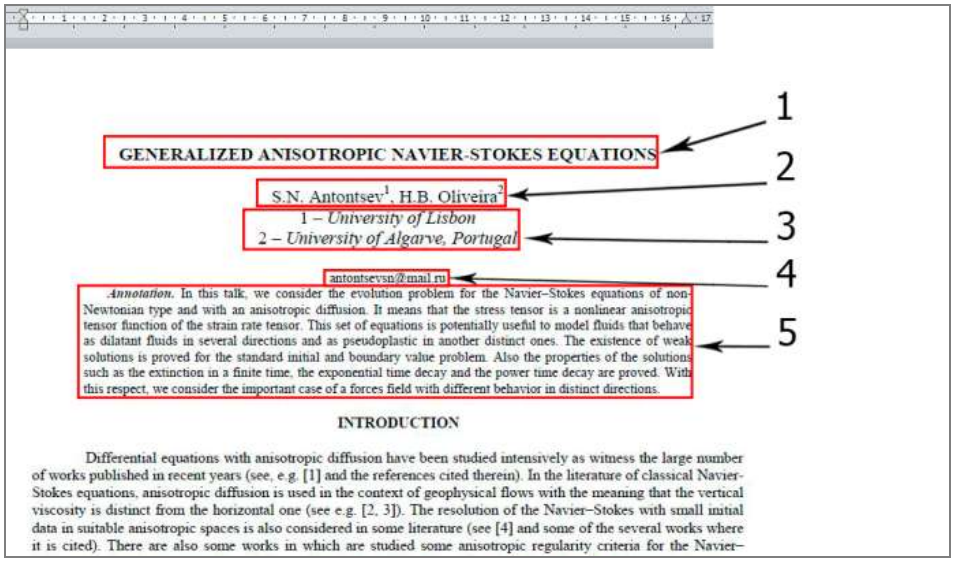

Figure 1. An example of structural analysis of text: 1 – article title, 2 – author list block, 3 – affiliations block, 4 – email block, 5 – annotation block. Stylish design corresponds to Table 1

VI. SERVICE FOR EXTRACTION OF BIBLIOGRAPHIC METADATA

The service workflow is implemented as following:

- from the original layout of the collection of proceedings of the Congress we extract a bibliographic description of each publication;
- we extract a block of literature; using regular expressions, we select bibliographic lists of articles; the distinguishing feature of these data is the presence of signs // in the bibliographic description; in this case, a regular expression was used in which individual blocks of metadata were allocated to groups;
- further we analyze the main metadata-we select authors, titles of articles, publications, etc.;
- using the developed web application to generate an XML-file, write it in accordance with the rules of the Russian Science Citation Index (RSCI); it contains a set of metadata for this publication;
- we upload a metadata set to the RSCI.

Thus, the created algorithm makes it possible not only to use semantic tools for working with digital content, but also to create automatically new types of documents.

REFERENCES

[1] S. Peroni. “Semantic Web Technologies and Legal Scholarly Publishing,” Springer International Publishing, 304 p, 2014, doi:10.1007/978-3-319-04777-5.

[2] S. Peroni. “Semantic Publishing: issues, solutions and new trends in scholarly publishing within the Semantic Web era,” Ph. D. Thesis. Department of Computer Science, University of Bologna, Italy, 221 p., 2012, http://amsdottorato.cib.unibo.it/4766/.

[3] J. Chen, H. Chen. “A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning,” Journal of Software, vol. 8, no. 1, pp. 55–62, 2013, doi:10.4304/jsw.8.1.55-62.

[4] E.V. Biryaltsev, A.M. Elizarov, N.G. Zhiltsov, E.K. Lipachev, O.A. Nevzorova, and V.D. Solovyev. “Methods for Analyzing Semantic Data of Electronic Collections in Mathematics,” Automatic Documentation and Mathematical Linguistics, vol. 48, no. 2, pp. 81–85, 2014, doi:10.3103/S000510551402006X.

[5] A.M. Elizarov, D.S. Zuev, E.K. Lipachev. “Mathematical Content Semantic Markup Methods and Open Scientific E-Journals Management Systems,” In: Klinov P., Mouromtsev D. (eds.) Knowledge Engineering and Semantic Web. Communications in Computer and Information Science, Springer, vol. 468, pp. 242–251, 2014, doi:10.1007/978-3-319-11716-4_22.

[6] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, F. Vitali. “The Document Components Ontology (DoCO),” Semantic Web, vol. 7, no. 2, pp. 167-181, 2016, doi:10.3233/SW-150177.

[7] A. Ruiz-Iniesta, and O. Corcho. “A review of ontologies for describing scholarly and scientific documents,” CEUR Workshop Proceedings, vol. 1155, pp. 1-12, 2014, http://ceur-ws.org/Vol-1155/paper-07.pdf.

[8] A.M. Elizarov, D.S. Zuev, E.K. Lipachev, and M.A. Malakhaltsev. “Services Structuring Mathematical Content and Integration of Digital Mathematical Collections at Scientific Information Space,” CEUR Workshop Proceedings, vol. 934, pp. 309–312, 2012, http://ceur-ws.org/Vol-934/paper47.pdf.

[9] A.M. Elizarov, E.K. Lipachev, Yu.E. Hohlov. “Semantic Mefhods of Structuring Mathematical Content Providing Enhanced Search Functionality,” Information Society, no. 1–2, pp. 83–92, 2013, http://elibrary.ru/download/elibrary_20376784_ 48362557.pdf.

[10] F. Ronzano, H. Saggion. “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. doi:10.1007/978-3-319-24282-8_18.

[11] D. Tkaczyk, B. Tarnawski and Ł. Bolikowski. “Structured Affiliations Extraction from Scientific Literature,” D-Lib Magazine, vol. 21, no. 11/12, 2015, doi:10.1045/november2015-tkaczyk.

[12] A.M. Elizarov, E.K. Lipachev, and S.M. Khaydarov. “Automated system of services for processing of large collections of scientific documents,” CEUR Workshop Proceedings, vol. 1752, pp. 58-64, 2016, http://ceur-ws.org/Vol-1752/paper10.pdf.

[13] A.M. Elizarov, A.V. Kirillovich, E.K. Lipachev, A.B. Zhizhchenko, and N.G. Zhil’tsov. “Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics,” Doklady Mathematics, vol. 93, no. 2, pp. 231–233, 2016, doi:10.1134/S1064562416020174.

[14] M.R. Kogalovsky, S.I. Parinov. “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, doi:10.1007/978-3-319-24543-0_7.

[15] O. Nevzorova, N. Zhiltsov, A. Kirillovich, and E. Lipachev. “OntoMathPRO Ontology: a Linked Data Hub for Mathematics,” In: Klinov P., Mouromtsev D. (eds.) Knowledge Engineering and Semantic Web. Communications in Computer and Information Science, Springer, vol. 468, pp. 105–119, 2014, doi: 10.1007/978-3-319-11716-4_9.

[16] A.M. Elizarov, E.K. Lipachev, O.A. Nevzorova, and V.D. Solov’ev. “Methods and Means for Semantic Structuring of Electronic Mathematical Documents,” Doklady Mathematics, vol. 90, no. 1, pp. 521–524, 2014, doi:10.1134/S1064562414050275.

[17] Standard ECMA-376: Office Open XML File Formats. http://www.ecmainternational.org/publications/standards/Ecma-376.htm.

[18] A.M. Elizarov, E.K. Lipachev, S.M. Khaydarov. “Semantic analysis of large scientific documents collections,” TEL-2016. Kazan: Kazan University Press. pp. 21-25, 2016, ISBN:978-5-00019-650-2, http://shelly.kpfu.ru/e-ksu/docs/F1866581513/1_7_PDF_proceedings__1_.pdf (In Russian).

[19] A.M. Elizarov, E.K. Lipachev, S.M. Khaydarov. “Automated system of structural and semantic processing of physical and mathematical content,” Uchenye Zapiski ISGZ, no. 1 (14), pp. 210–215, 2016, http://kpfu.ru//staff_files/F598686152/Elizarov_Lipachev_Khaidarov.pdf (In Russian).

[20] S.M. Khaydarov. “Semantic analysis of documents in the control system of digital scientific collections,” Russian Digital Libraries Journal, vol. 18 (1-2), pp. 61-85, 2014, http://ojs.kpfu.ru/index.php/elbib/article/view/19 (In Russian).

[21] Document Object Model (DOM). https://www.w3.org/DOM/.