Developing a Machine-readable Catalog of Computer Programs and Tools for Extracting and Analyzing Contextual Knowledge

Olga Kononova [1][0000-0001-6293-7243] and Dmitry Prokudin[1,2][0000-0002-9464-8371]

[1] ITMO University, 49 Kronverksky, 197101 St Petersburg, Russia
[2] St. Petersburg State University, 7/9 Universitetskaya emb., 199034 St Petersburg, Russia
kononolg@yandex.ru, hogben.young@gmail.com

Abstract. For researchers in modern conditions of development and the total application of information and communication technologies, there is an issue of choosing effective tools for research purposes. A huge amount of the existing software lacks classifications of the software and information systems to consider research task classes. The project implemented by the authors aimed to develop an approach to research the evolution of the thematic and terminological apparatus of interdisciplinary scientific fields. The following methods: search, extraction, clarification, explication, analysis, and presentation of contextual knowledgewith software and information systems were considered and applied. The specifics of the research limits software and information systems to the tasks of contextual scientific knowledge processing. The main types of software and information systems used for these purposes were analyzed, and their main functional characteristics were identified. Based on the typology of contexts and the groups of characteristics identified, an approach is proposed to develop a catalog of software and information systems analysis of contextual knowledge with the functions of allocation, classification, and explication of scientific content. To provide information about software and information systems in the catalog a Dublin Core metadata model is proposed. This model allows not only to describe and structure the main characteristics of software and information systems, but also to present the catalog in a machine-readable form to add new records, efficiently search the necessary software and information systems subject to the research tasks, and integrate it into the scientific information network based on open science principles. A palliative solution for testing the correctness of Dublin Core metadata presentation and metadata exchange via the OAI-PMH Protocol is presented.

Keywords: Software, Classification, Catalog, Contextual Knowledge, Dublin Core, OAI-PMH.

Introduction

The variety of software, application environments, and web-oriented services for various purposes makes it difficult for modern researchers to choose tools that can be effectively used in scientific research. They have to focus on the existing approaches

to software classification. Both general approaches that distinguish the common software classes and special approaches that focus on a more detailed description of software subclasses for various applications have been developed. The most common classifications include, for example, the Classifier of programs for computers and databases used by government agencies in Russia [24]. One of the general approaches to classify the Business intelligence (BI) class software is the annually updated analytical report "Infrastructure and Applications Worldwide Software Market Definitions" by Gartner [12], which reflects the general approach of Gartner to assess the software market development [32]. This approach is followed by the International Data Corporation (IDC), one of the world's largest consulting companies, a leading provider of information and consulting services, and an organizer of events in the information technology, telecommunications, and consumer technology markets [2]. The Computing Classification System (developed by the Association for Computing Machinery, the latest version was introduced in 2012) can also be used to classify software and information systems. The system is presented as a single source of categories and concepts that reflect the current state of the fields related to computational engineering, computer science, and information and communication technologies [5]. Other approaches to classification, which have historically been formed from the logic of computer technology and software development, are also widespread [15, 16, 22].

The analysis of approaches to software classification makes it possible to identify the problem of choosing a specific tool with which researchers, as users of software products, can solve research and analytical tasks. This problem arises from the fact that software developers try to integrate the maximum set of functionalities into their systems and thus ensure that a whole range of tasks is performed. Unfortunately, the results obtained are often not equivalent. A specific computer application has its own "specialization", that is, a minimal set of functionalities that ensures the performance of a narrower range of tasks, but most optimally and successfully. There are usually many analogs, their choice by a researcher is based on compliance with certain initial requirements to the system (for example, a particular language support or implementation of certain methods in the software). There are information systems that were not designed by the developers to process scientific information or analyze information for scientific purposes. Although, if there is an introductory, explanatory information or a developed method, they can be efficient for scientific purposes. Even with detailed information about a specific computer program or an information system, it is not always possible to categorize it with the existing classifications. Therefore, both classification and identification of specific functionality can be made or refined only in the process of software application to solve specific research tasks. The presence of analogs initiates creation of software catalogs focused on solving a certain class of research tasks based on the classification developed.

1 Approach to software search and selection

As part of the ongoing project to develop an approach (synthetic method) to In the ongoing project framework,the principle of the synthetic method independence from the specific tools was chosen as the main one.

The research specifics led to the development of a typology of contextual knowledge of textual modality, which must be considered when choosing an appropriate software for processing contexts of certain types (corpus, fragment, paragraph, sentence, term-concept, thesaurus, meta-description, semantic group, thematic collection). The topology allows software application for multi-level structural analysis that increases the research task effectiveness. Focusing on the methods used in the study (search, extraction, explication, analysis, and representation of contextual knowledge), the general classifications revealed a lack in clear division into classes for these methods as enlarged functions. It is also not possible to determine the types of contexts being processed using these classifications.

For example, Gartner identifies the following market segments:

- Data Warehouse;
- On-Line Analytical Processing, OLAP;
- Enterprise Information Systems, EIS and Decision Support Systems, DSS;
- Data Mining;
- Query and Reporting Tools.

In the Computing Classification System, the following subclasses correspond to the systems considered:

- Specialized information retrieval – Structure and multilingual text search;
- Document management and text processing – Document capture – Document searching; Document analysis.

In the Russian Classifier, the following subclasses of software application can be associated to the programs under consideration:

- Search engines – software systems that search for text, graphics, and other information in local, corporate, and other repositories, including consulting and information systems for searching and viewing information in specialized multi-industry databases;
- Linguistic software – parsers and semantic analyzers/systems for natural language text analysis with the selection of syntactic sentence structures or semantic relations between text elements and general text meaning;
- Systems for data sets collection, storage, processing, analyzing, modeling and visualizing – business analysis systems (BI)/programs focused on big unstructured data processing to facilitate their interpretation, including tools for data extraction and transformation (ETL), subject-oriented information databases (EDW), tools for real-time analytical processing (OLAP), data mining, generating reports, graphs, charts and other visual forms, decision support (DSS).

To search and identify a software with the functions of extracting, classifying, and explicating scientific content to support scientific research, both open network sources and scientific publications with similar software were used. The analysis of the identified systems showed that the vast majority are used as text linguistic analysis tools

in linguistics [13, 33], sociology [31], cybersecurity [21], as well as in interdisciplinary fields [4, 10, 18].

Some linguistic software catalogs are presented in the Internet [1, 14, 18, 22, 28, 29, 30]. This software can be used for research purposes.

Catalog analysis allows to identify common areas of software used for the tasks listed above:

- Text Mining;
- Text Analytics/Analysis;
- Information Retrieval/Extraction;
- Text Comparison;
- Topic Clustering/Modelling;
- Text Visualization.

Such catalogs do not have classifiers that consider their main functional purpose and the types of contexts being processed. Therefore, they do not perform the tasks of selecting effective tools for research. It was decided to reject the use of other software classifications widely presented in the network, including those based on the signs "scope of application" and "system functionality", for the same reasons. Such classifications are intended for the business community, operate with business concepts, focus on the range of business tasks of an enterprise, organization, or market analysis.

In this regard, the purpose of this study is to develop a structured description of a software using the main characteristics: software class, main functions, and types of the processed contexts. Based on a structured description, the software catalog development is performed, which allows researchers to solve the problem of effective software selection to meet the research aims and objectives. The creation of the catalog solves the researcher's pragmatic task to make an informed choice of a required software set.

2 Development of a catalog of a software with functions and services for extracting and analyzing contextual knowledge for scientific research

2.1 Defining the main software classes

When forming the structure of a software description for the developed catalog, the general classification was chosen based on the this software applicability for the analysis of contextual knowledge with the functions of extracting, classifying, and explicating scientific content to support scientific research.

The target group of catalog users includes scientists and teachers specializing in interdisciplinary research and working with various sources of information and big data. The following specific task classes for the interdisciplinary studies were set:

- Neural Network;
- Machine Learning;
- Natural Language Processing;

- Information Extraction;
- Ontology;
- Forecasting Systems;
- Creation and Use of Thesauri;
- Topic Clustering/Modelling;
- Full-text Databases;
- Abstract Databases (metadata only).

These classes of tasks in the catalog correspond to the "type of software" characteristic. The latter two types are characterized only by full-text and abstract databases that have their engines for searching, selecting, and analyzing information.

The integrated types of software do not always reflect the diversity of their functional capabilities. Therefore, for a more complete understanding of their capabilities and rational choice for specific research purposes, the main functions of the software are grouped separately. The following main functions of the software are distinguished:

- Classification;
- Forecasting;
- Contextual analysis;
- Selection of data according to various criteria (smart search);
- Automated data/metadata exchange;
- Visualization.

This is a set of basic functions. However, when a specific software application is included in the catalog, its analysis may reveal other specific functions. Therefore, the classification of the functions is extensible.

The proposed classification does not consider some important classes of tasks, such as "scientific communication" and "issues of science management and research coordination", as they go beyond research tasks.

2.2 A typology of contexts for software features

The choice of specific software for research purposes is related to its ability to process certain types of contexts. Within the framework of this research, the concept of context is understood as an independent conceptual unit of the thesauri, used as a basis for classifying scientific texts, as well as for visualizing hierarchical and associative relations between terms. The explication and analysis of contextual knowledge resulted in a typology of contextual knowledge developed in this project [17].

This classification can be further specified for the study of more specific subject areas. The correlation of the software with the types of the processed contexts also allows researchers to choose a software more rationally. Therefore, when classifying a software, it is proposed to use the type of the processed contexts as an essential characteristic.

Based on the types of stored, extracted, and processed (analyzed) contexts and their specification, the software for the catalog can be divided into the following enlarged categories:

- An information search system that processes a large number of unstructured texts and multimedia information, with limitations in the user dialogue with the system, as well as limited graphematic analysis and low reliability of link detection (Yandex, Google);
- Information systems that represent text databases, digital online archives of scientific publications and abstract databases of multidisciplinary areas, that significantly differ in the content analysis functionality (eLibrary, T-Libra, Science Direct, Scopus and WoS);
- Information and analytical systems that have various degrees of completeness of fact detection and self-learning, levels of semantic hierarchy and automatic logical analysis of factual information. These systems also process a large volume of unformalized texts and multimedia information, (Mallet, AskNet, Voyant-Tools, Tropes, Sketch Engine, CLAVIRE, RCO (Russian Context Optimizer));
- Multifunctional mixed-type information systems that have the advantages and disadvantages of the information systems described above: the ability to process a large volume of unformalized texts and multimedia information, reliability of identifying links, a wide range of document formats. These systems have limitations in automatic semantic analysis of various levels, which is a software developer task (ABBYY Intelligent Tagger SDK, ABBYY Smart Classifier SDK; Title: PROMT Analyser).

These enlarged categories are used in the catalog to group software.

2.3 Machine-readable view of the catalog

The description of the software for contextual knowledge processing was analyzed. This leads to the conclusion that such catalogues are mostly static lists or tables, where the software is either grouped by a certain attribute (for example, freely distributed or commercial; belonging to enlarged functional categories) or presented in an unstructured form with a brief description of features and links to relevant sites on the Internet. This presentation of information makes it difficult to quickly search for and effectively select the software necessary for conducting research, considering the main features, functionality, types, and formats of the processed contexts.

For a structured representation of information about the software, it is suggested to use its description, as well as the descriptions of the documents, via the metadata representation. For example, González and van der Meer considered standard metadata representation formats (Dublin Core, EAD, ISAD (G) and MARC) and suggested the Extended Dublin Core for Software Components (XDC-SC) of the Dublin Core scheme, which allows extracting information about a software using standard search engine tools or XML tools [11]. They also suggested that this approach may encourage the creation of environments to present information about a software. Other researchers suggest not to focus on one standard, but develop a Semantic Master Metadata Catalog (SMMC) to ensure interaction between the existing metadata models (such as Dublin Core, UNIMARC, MARC21, RDF/RDA, and BIBFRAME) based on the ontology mapping model [20]. Here, an approach to developing a Semantic enhanced Metadata

Software Ecosystem (SMESE) is proposed. This ecosystem is designed to support specific distributed content management applications. However, the implementation of such a solution is a complex task that can only be solved at a large consortium level.

Based on the generally accepted approaches for software description, it is proposed to use the Dublin Core meta-data representation specification in the catalog. The main characteristics of the presented software are described by the corresponding elements of the main metadata set (Dublin Core Metadata Element Set, DCMES) [6]. In this approach, the combination of element values sufficiently describes the software presented in the catalog, in accordance with the general approach of this specification application to describe various entities [9, 23, 27]. The proposed approach also makes it possible to present the catalog in a machine-readable form in the network information systems with free access for both researchers and automated search and identification by search engines. In this case, the users can search for various metadata elements: classes, software functions, or types of the processed contexts.

Usually, this approach mainly describes various text objects: articles, books, library catalog cards, archive materials, and semantic models [27]. When developing a metadata representation scheme for software description, its specifics is not considered, which is important when choosing a software (for example, the types of contextual knowledge being processed) [3, 11]. Therefore, in connection with the specifics of the software description, in addition to the main metadata elements, qualifiers were used to refine the characteristics, which make the second level of metadata and refine the elements [7, 20].

Also, the most suitable software platform was selected to present the catalog in a machine-readable form. When choosing, the following basic principles were considered:

- availability – open source or non-commercial software;
- popularity – the most well-known and widespread solution;
- flexibility – ability to adjust the metadata description to the task of catalog creation.

Among the systems considered, the most popular at present is DSpace software platform (https://duraspace.org/dspace/). According to the most authoritative aggregator ROAR [26], of the 4,725 open access repositories registered, 1,965 use DSpace. The second place takes EPrints (679). DSpace meets all the above criteria, so this platform was chosen for this study. After reviewing the documentation for metadata presentation in DSpace by the Dublin Core specification [8], and considering the specifics of qualifiers application, the following set of metadata is proposed for each catalog record:

`dc.title` — software title ;
`dc.creator` — developer;
`dc.subject.classification` — main functions (can be added after analyzing the corresponding software);
`dc.subject.other` — type of context to process;
`dc.description.abstract` — software description ;
`dc.publisher` — vendor (copyright holder);

dc.contributor — contributor (people or organizations who also participated in the software development);
dc.date.issued —last release date (year);
dc.type — categories (software classes according to the developed classification);
dc.format.mimetype — formats of the processed files;
dc.identifier.uri — identifier (link on the Internet to the developer's site);
dc.source.uri — source (link to the web application);
dc.language — languages of documents to be processed;
dc.relation.isreferencedby — relations (list of publications on the software);
dc.coverage — supported operating systems;
dc.rights.license — license type.

Based on the proposed approach, more than 50 software items were described. A visual representation of a catalog record by the Dublin Core specification is shown in Figure 1.

dc.title	Voyant-Tools
dc.creator	Stéfan Sinclair, McGill University; Geoffrey Rockwell, University of Alberta
dc.subject.classification	processing of individual documents; processing by a collection of documents (text corpora); classification; analysis of Internet pages; frequency analysis; context analysis; trend contextualization (building trends); data analysis visualization
dc.subject.other	term; paragraph; document; document collection
dc.description	Web-based system for downloading and analyzing digital texts, studying the frequency and distribution of terms in documents, and in a collection of documents (corpora). It is a set of different functional modules. There is a local solution in the form of an application on JETTY
dc.publisher	Voyant-Tools
dc.contributor	Andrew MacDonald; Cyril Briquet; Lisa Goddard; Mark Turcato
dc.date.issued	2018
dc.type	natural language processing
dc.format.mimetype	txt; rtf; doc; docx; pdf; zip; html; xml
dc.identifier.uri	https://voyant-tools.org
dc.source.uri	http://voyeurtools.org/voyant-server/
dc.language	Multilanguage
dc.coverage	Web-based application (Web interface), Mac, Windows, JETTY server, Voyant server
dc.rights	Freeware software
dc.relation.isreferencedby	Laurie J. Sampsel (2018) Voyant Tools, Music Reference Services Quarterly, 21:3, 153-157, DOI: 10.1080/10588167.2018.1496754; Sinclair, Stéfan; Rockwell, Geoffrey (2016). "Voyant Facts". Hermeneutica: Computer-Assisted Interpretation in the Humanities. Stéfan Sinclair & Geoffrey Rockwell. Retrieved 2016-12-20; Using Voyant for Text Analysis: http://voyeurtools.org/using-voyant-for-text-analysis/;

	Rambsy, Kenton (2016). Text-Mining Short Fiction by Zora Neale Hurston and Richard Wright. Using Voyant Tools // CLA Journal. № 59 (3): 251–258; Priestley Alexis. Voyant Tools: A Tutorial for Text Analysis: https://medium.com/@priestleyal/voyant-tools-a-tutorial-for-text-analysis-df265d85d214;
dc.coverage	Multisystem
dc.rights.license	Creative Commons Attribution 4.0 International (CC BY 4.0)

Fig. 1. Description of Voyant-Tools by Dublin Core specification

2.4 The implementation and use of the catalog

Despite the choice of the DSpace software platform for machine implementation, this system installation and configuration is not a trivial task, which did not allow to implement this solution immediately. In this regard, a free and open-source software Open Journal Systems (OJS, https://pkp.sfu.ca/ojs/) was chosen as a palliative solution for initial testing. This system is a full-cycle publishing platform to publish electronic journals. This solution has already been applied to the machine-readable representation of the thesaurus in the framework of the ongoing project. The OJS is easier to install and configure and works on most virtual hosts. This system has all the functionality required: it supports Dublin Core metadata format, allows to search for metadata, provides open access to information, and acts as a provider for OAI-PMH protocol. For experimental purposes in the installed OJS system (http://ojs.iculture.spb.ru/index.php/thesauri), the descriptions of several software units from the selected ones were entered. The Open Harvester Systems (OHS, https://pkp.sfu.ca/ohs/) installation was used to control the correctness of metadata display and verify the operation of OAI-PMH Protocol., which is an OAI-PMH metadata aggregator. The proposed approach to present the software catalog in the machine-readable form also allows to export metadata to other presentation formats for integration in various information systems and metadata aggregators. Using the OAI-PMH metadata exchange protocol makes it possible to integrate the catalog into the information space of scientific research. The researchers can not only search the catalog but also create their information systems and aggregate information from the catalog. A platform like DSpace also allows to use it as an aggregator and collect information about the software application presented in the catalog from various resources using OAI-PMH protocol. For example, these may be scientific publications that consider particular software for specific scientific purposes.

The implementation of such a distributed information environment allows the users to present not only software descriptions but also information about its application in one information space. This provides the researchers with methodological support for a more rational choice and tools efficiency for their scientific purposes.

Conclusion

The research has shown that there is no common approach to classifying the software designed for analyzing contextual knowledge with the functions of highlighting, classifying, and explicating scientific content, considering the types of the processed contexts. It was also found that there are no developments in the representation of software catalogs in machine-readable form based on metadata format and considering the specifics of software for contextual knowledge analysis.

The developed approach to present a catalog of a software designed for contextual knowledge analysis with the functions for highlighting, classifying, and explicating scientific content based on Dublin Core provides:

- integration of the developed context typology into the catalog, which is an essential characteristic and the basis for choosing a software for conducting specific research;
- creation of a machine-readable catalog with standard freeware software (for example, OJS, DSpace);
- efficient search and selection of a specific software for research purposes by the main characteristics described in Dublin Core tags, using standard search engines;
- open access to catalog records for both users and automated indexing;
- automated exchange over OAI-PMH protocol for aggregation of catalog meta descriptions in other information systems.

The proposed palliative solution (OJS) is expected to be replaced in the future with an information system based on the freeware DSpace. In parallel, the work will continue filling the catalog with software descriptions that can be used for contextual knowledge analysis with the functions for extracting, classifying, and explicating scientific content. Scientific publications describing the research results using the software presented in the catalog will also be selected.

There is a great potential to further catalog application in the framework of teaching activities: masters of "Digital smart city technologies" educational program, majoring in "Applied Informatics" will use it to select the technological tools for their research projects. The catalog is one of the components of the educational and methodological complex "Technologies of data extraction and mining in scientific research", aimed at forming research and analytical competencies of undergraduate students. The course "Information technologies in science" will be modified based on the developed educational and methodological complex.

Acknowledgement. This work was supported by the Russian Foundation for Basic Research (project #18-011-00923-a) and the Vladimir Potanin Foundation (project GK200000654).

References

1. 4 Free and Open Source Text Analysis Software, https://www.softwareadvice.com/resources/easiest-to-use-free-and-open-source-text-analysis-software, last accessed 2020/02/17.

2. Andsbjerg, R., Vesset, D.: IDC's Worldwide Software Taxonomy, 2018: Update, https://www.idc.com/getdoc.jsp?containerId=US44835319, accessed 2020/02/17.
3. Brisebois, R., Abran, A., Nadembega, A.: A Semantic Metadata Enrichment Software Ecosystem (SMESE) Based on a Multi-Platform Metadata Model for Digital Libraries. Journal of Software Engineering and Applications. 10, 370-405 (2017). DOI: 10.4236/jsea.2017.104022.
4. Chugunov, A.V., Kabanov, Y.: "Electronic Governance" As an Interdisciplinary Scientific Field: Scientometrics Analysis. In: The State and Citizens in the Electronic Environment. Vol. 3. Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019, pp. 11 – 24 (2019). DOI: 10.17586/2541-979X-2019-3-11-24 [in Russian].
5. Computing Classification System, https://dl.acm.org/ccs, accessed 2020/02/17.
6. DCMI Metadata Terms. Dublin Core Metadata Initiative, https://www.dublincore.org/specifications/dublin-core/dcmi-terms, accessed 2020/02/17.
7. DCMI Qualifiers. Dublin Core Metadata Initiative, https://www.dublincore.org/specifications/dublin-core/dcmes-qualifiers, accessed 2020/02/17.
8. DSpace/dublin-core-types.xml at master DSpace. DSpace. GitHub, https://github.com/DSpace/DSpace/blob/master/dspace/config/registries/dublin-core-types.xml, accessed 2020/02/17.
9. Fedotov, A.M., Leonova, Y.V.: Requirements for the prototype of the information resources management system in distributed information systems for the support of scientific research. Computational technologies. 23(5), 82-109 (2018). DOI: 10.25743/ICT.2018.23.5.008 [in Russian].
10. Geger, A.E., Tchupakhina, Y.A., Geger, S.A.: Computers programs for the qualitative and mixed data analysis. St. Petersburg Sociology Today. 6, 374-388 (2015).
11. González, R., Van Der Meer, K.: Standard Metadata Applied to Software Retrieval. Journal of Information Science. 30(4), 300–309 (2004). DOI: 10.1177/0165551504045850.
12. Infrastructure and Applications Worldwide Software Market Definitions. Gartner Dataquest Guide (2002), http://smartshore.us/Infrastructure_Market_trends_2003.pdf, accessed 2020/02/17.
13. Ivanova, A.A.: Rhetoric of wargames (results of the content analysis). In: Computer Linguistics and Computing Ontologies. Vol. 3 (Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019), pp. 266 – 278. ITMO University, St. Petersburg (2019). DOI: 10.17586/2541-9781-2019-3-266-278 [in Russian].
14. Catalog of linguistic programs and resources on the Web. Compiled by S.V. Logichev. (2006), https://rvb.ru/soft/catalogue/index.html last accessed 2020/02/17. [in Russian].
15. Classification of computer software. In: Samsonova, O.V. Informatika: uchebnoe posobie, http://tpt.tom.ru/umk/informat/uchebnik/klass.htm, last accessed 2020/02/17. [in Russian].
16. Classification of software. In: Alekseev, E.G., Bogatyrev, S.D. Informatika. Mul'timediynyy elektronnyy uchebnik, http://inf.e-alekseev.ru/text/Klassif_po.html, last accessed 2020/02/17. [in Russian].
17. Kononova, O.V., Prokudin, D.E.: An approach to extraction, explication and presentation of contextual knowledge in the study of developing interdisciplinary research areas. International Journal of Open Information Technologies 8(1), 90-101 (2020). [in Russian].
18. Kravchenko, Yu.A.: Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems. Izvestiya SFedU. engineering sciences 7(180), 5-18 (2016). DOI: 10.18522/2311-3103-2016-7-518 [in Russian].

19. Kuznetsov, K.I.: Overview of systems for extracting data from unstructured texts (2013), http://www.pullenti.ru/(X(1)S(ngdeikpifqat0ccmnoqanfz3))/CompetitorPage.aspx?AspxAutoDetectCookieSupport=1, last accessed 2020/02/17. [in Russian].
20. Dublin Core Qualifiers. RUSMARC, Russian version of UNIMARC. National Library of Russia, http://www.rusmarc.info/soft/dcq.html, last accessed 2020/02/17. [in Russian].
21. Lavrent'ev, A.M., Smirnov, I.V., Solov'ev, F.N., Suvorova, M.I., Fokina, A.I., Chepovskiy, A.M.: Analysis of corpus of extremist texts and unlawful texts. Voprosy kiberbezopasnosti 4(32), 54-60 (2019). DOI: 10.21681/2311-3456-2019-4-54-60.
22. Morozevich A.N. et al.: Fundamentals of Informatics and Computer Engineering: A Study Guide. Morozevicha, A.N. (eds). BGEU, Minsk (2005). [in Russian].
23. Noor, S., Shah, L., Adil, M., Gohar N., Saman, G.E., Jamil, S., Qayum, F.: Modeling and representation of built cultural heritage data using semantic web technologies and building information model. Computational and Mathematical Organization Theory 25, 247–270 (2019). DOI: 10.1007/s10588-018-09285-y.
24. Order of the Ministry of Telecom and Mass Communications of the Russian Federation of December 31, 2015 N 621 "On approval of the classifier of programs for electronic computers and databases" (ed. Order of the Ministry of Telecom and Mass Communications of 01.04.2016 N 134, of 30.07.2019 N 422), https://normativ.kontur.ru/document?moduleId=1&documentId=345157#h74, last accessed 2020/02/17. [in Russian].
25. Programs for linguistic analysis and text processing, http://asknet.ru/analytics/programms.htm, accessed 2020/02/17. [in Russian].
26. Registry of Open Access Repositories, http://roar.eprints.org, accessed 2020/02/17.
27. SUNScholar/Metadata/By Function. Libopedia, https://wiki.lib.sun.ac.za/index.php/SUNScholar/ Metadata/By_Function, accessed 2020/02/17.
28. Text Analysis, Text Mining, and Information Retrieval Software, https://www.kdnuggets.com/software/text.html, accessed 2020/02/17.
29. Text Mining Software, https://www.capterra.com/text-mining-software, accessed 2020/02/17.
30. Text mining, text analytics & content analysis with free open source software, https://www.opensemanticsearch.org/doc/analytics/textmining, accessed 2020/02/17.
31. Vidiasova, L., Tensina, I.: Results of the Semantic Analysis of Texts in Mass Media on the Development of «Smart Cities» in Russia. In: The State and Citizens in the Electronic Environment. Vol. 2 (Proceedings of the XXI International Joint Scientific Conference. Internet and Modern Society, IMS-2018, St. Petersburg, May 20 - June 2, 2018), pp. 112-117. ITMO University, St. Petersburg (2018). DOI: 10.17586/2541-979X-2018-2-112-117.
32. Woodward, A., Anderson, R., Biscotti, F., Contu, R., Gupta, N., Hunter, E., Hare, J., Bhullar, B., Dayley, A., Roth, C., Swinehart, H., Dsilva, V., Wurster, L., Poulter, J., Palanca, T., Deshpande, S., Pang, C., Abbabatulla, B., Warrilow, M., Dharmasthira, Y., Kostoulas, J.: Market Definitions and Methodology: Software (2019), https://www.gartner.com/en/documents/3906823/market-definitions-and-methodology-software, accessed 2020/02/17.
33. Zhang, P., Zakharov. V. P.: Computerized visualization of the Russian language picture of the world. In: Computer Linguistics and Computing Ontologies. Vol. 3. Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019, pp. 92 – 105 (2019). DOI: 10.17586/2541-9781-2019-3-92-105.