Olga V. Kononova

Associate professor, PhD (Economy), ITMO University,
Saint-Petersburg, Russia
e-mail: kononolg@yandex.ru

Dmitry E. Prokudin

Associate professor, Dr. Science (Philosophy),
Saint-Petersburg State University, ITMO University,
Saint-Petersburg, Russia
e-mail: hogben.young@gmail.com

Elena E. Yelkina

Associate professor, PhD (Philosophy), ITMO University,
Saint-Petersburg, Russia
e-mail: e.e.e.1@mail.ru

УДК: 004:89+338.1
DOI: 10.24411/2079-0910-2019-13006

Contextual Knowledge Extraction: Terminological Landscape of Digital Economy

The article is based on the report made by the authors on the “International Conference on Materials, Applied Physics & Engineering (ICMAE-2018)” in Madhya Pradesh, India during 3–4th June, 2018. The aim of the research is to demonstrate the effectiveness of the synthetic method for contextual knowledge extraction from distributed information network environment for studying new trends of scientific directions. The actual objective is to reveal the trends in the corpus formation of ‘Digital Economy’ as a developing interdisciplinary research direction. In the present study, we offer the synthetic method, which combines various approaches and tools of Digital Humanities. This method is used to meet the challenges of the digital information resources selection, explication of the context knowledge and clarification of the corpus of emerging interdisciplinary scientific direction ‘Digital Economy: e-governance and smart technology’. The base of the research has proved possibility of synthetic method application to the study of trends in the development of thesauri of interdisciplinary scientific fields. The results of the research can be used to develop interdisciplinary ontology of ‘Digital Economy’.

Keywords: Digital Economy, e-Governance, contextual knowledge, context, synthetic method, smart technology, distributed network environment, information resources, interdisciplinary scientific directions.

Acknowledgment

The research was carried out with support from the Russian Foundation for Basic Research (RFBR) according to the research grant No. 18-011-00923 (“The working out of an integrated approach to the analysis of the developing terminological base of interdisciplinary areas of research in distributed network environment”).

Introduction

Global information society is changing its requirements for the quality of the information search as well as retrieval systems, including extraction methods and analysis tools, used for scientific research. Main requirements in the last case are not only transparency and reliability of data, but also ability to manage information resources independently (search, analyze, evaluate information) and thereby to obtain new expertise. The satisfying of information needs of a society increases management efficiency of enterprises, organizations, public sector of the economy as well as public institutions, all types and aspects of human activity.

The goal of modern interdisciplinary studies is analysis of perspective interdisciplinary scientific trends, forecast demand for their results in various areas of public life. One of the main trends of modern interdisciplinary studies is explication of contextual knowledge through the application of methods, approaches, technologies and tools of Digital Humanities. They include the following ones:

— methods and technologies of contextual knowledge extraction from large amounts of data — Data Mining and Big Data;

— methods and research technologies of thematic and semantic contexts automatically extracted from unstructured sources (contextual search). Contextual search presupposes the existence of structured sets of sources and a clear understanding of the subject.

Application of Digital Humanities in interdisciplinary studies is conditioned by objective factors connected with the development of modern interdisciplinary scientific directions:

— the steady increase of information volume generated within above mentioned directions;

— non-formalization and heterogeneity of the information;

— distribution of its storage and access;

— redundancy and polysemy of the corpus used. The corpus of interdisciplinary scientific directions is formed both through terms transferring from one subject domain to another without any necessary interpretation and adaptation, and through direct borrowing from foreign scientific sources.

Drawing Digital Humanities research technologies and tools in interdisciplinary studies is also determined by the broad development of network and distributed access to the information resources (scientific data and knowledge), and high speed of knowledge up-to-date. These days we can observe lagging behind the development of science-metric disciplines from the growth of interdisciplinary scientific corpus, which is generated unmanageably by research schools, groups, individual researchers. Polysemy of terminology and non-structured information even with free access to it does not allow researchers to track emerging trends and relationships efficiently. This leads to the loss of an important part of scientific knowledge and hypotheses that did not receive immediate wide spreading. Such knowledge was subsequently forgotten or thrown away as unclaimed one. The above-mentioned content determines the actuality of the solution to the problem of updating corpus of developing interdisciplinary research involving methods and tools of Digital Humanities.

In order to evaluate the dynamics of research interest to subject domain ‘Digital Economy’ in such aspects as e-governance, e-government, digital technologies as well as state services the study proposes a review of the corpus of Russian scientific journals. Discourse analysis can be seen as the theoretical basis of the study. It means that the language of scientific publications reflects the research directions, focuses on a number of scientifically significant results as well as allows forecasting of perspective study directions.

The review of methods, technologies and tools

In modern dynamically developing information society the basic mechanism of information volume increase is moving to its digital representation. The use of various technologies and tools of textual information representation in digital form leads to acceleration processes of generating information and increase in volume of electronically stored information. Information and communication technologies are mostly used to access such information. The development of communication technology has resulted in global and broad access to accumulated data of digital information. It is fully applied to scientific knowledge, which is now represented in digital forms [*Borgman,* 2007; *Borgman.,* 2009; *Weller,* 2011].

Technologies aimed at scientific knowledge accumulating and providing on-line access to it are the most developed ones. Digital scientific resources generally have a powerful search engine that allows searching the necessary information according to specified criteria quickly and efficiently. Access to data arrays distributed in digital scientific Internet resources provides a solution to the problems of availability and rapid dissemination of research results [*Laakso,* 2014; *Scanlon,* 2014]. Technologies of digital scientific resources give a possibility to solve the problem of long-term storage of scientific information, [*Agosti et al.,* 2003; *Burda, Teuteberg,* 2013; *Nielsen, Hjørland,* 2014]. Text sources in modern studies extracted from the digital scientific resources are more often used. Study of semantic and thematic contexts taken from those sources is important for tasks of information technology use in science and education, management and business. Contextual knowledge (i.e. knowledge contained in various contexts, derived from the full-text queries) can exist in various types and forms; can be extracted and studied as a result of the full-text search and its subsequent processing. In the world, scientific discourse study of contextual knowledge is carried out primarily in two directions: the theory and practice of content analysis; the theory and practice of context learning.

Study of contexts of different types, forms and levels are carried out in the framework of the traditional content analysis: the method and technology of qualitative-quantitative analysis of documents in order to identify or measure social facts and trends reflected in those documents. Content analysis examines the documents in the social context. Content analysis of media reports based on the paradigmatic approach is increasingly distributed. According to this approach studied signs of texts (content problems, causes, the problem-formatted subject, the degree of problem tension, the ways of its solution, etc.) are regarded as a definite organized structure.

The tasks study of contextual knowledge is fulfilled through the following approaches:

1. Technological approach is aimed at the developing of information systems and implementing contextual search algorithms in search engines [*Osipov et al.,* 2004; *Tao et al.,* 2013].
2. Semantic approach is aimed at the developing and using of linguistic methods to the analysis of tests and identification of the certain meanings in those tests [*Chernij, Tuzovskij,* 2009; *Lyapin et al.,* 2016; *Turney et al.,* 2010].

3. Content approach means an applied use of information search and analysis algorithms in functioning information systems for quantitative and qualitative analysis of their content processing from the definite subject domains [*Saifa et al.,* 2016].

Moreover, one more important research direction is the analysis of heterogeneous data processing capabilities (texts), distributed in heterogeneous information systems with access from both global and local networks. This field is aimed at implementing Grid technologies information distribution to its systematization and integrity at its constant quantitative augmenting [*Ageev et al.,* 2002; *Borodkin,* 2008; *Zhizhimov et al.,* 2011].

These days quite a lot of software products targeted at providing specialized services for handling unstructured textual and non-textual information are worked out. Basically, those are the programs of linguistic analysis of the text. They are designed for special tasks for text analysis, but not all of them provide full text search. Their list and detailed description is presented on: http://asknet.ru/analytics/programms.htm. Analysis of their capacity allows concluding that retrieval systems of Yandex Server and Nigma are the closest on functional to the realization of the present study objectives.

Among the leading study centers of contextual knowledge, methods and technologies of its extraction and analysis we can name university and specialized laboratories and research centers, and large commercial organizations. The main research projects, implemented in the last 10 years, include the following:

1. The Sociological Institute of the Russian Academy of Sciences (St. Petersburg, Russia): Context-oriented Methods to Build Social Knowledge. Context-oriented Ontologies of Sociological Research (2009–2011). Working out Constructive Methods to Develop Sociological Knowledge (2012–2014). Context-oriented Methods to Construct Theories (2006–2008). Graph context-oriented ontological methods of knowledge management are used to solve problems of operating with sociological concepts and explicit relationships between them by creating relationships of computer functionality (inheritance, encapsulation, type-creating, etc.). Keeping preterition, checking the concepts polysemy, maintaining their identity being used in different contexts; analytical estimating methods of qualitative research thesaurus is supported. The research package of applied computer programs, used in listed research, program-simulating the proposed methods, is not an intuitive clear tool. Therefore, it requires prior preparation. This approach is useful for performing large scale modeling projects of subject domain knowledge.
2. The Higher School of Economics (Saint-Petersburg, Russia): working out the concept and methodology of multilevel monitoring the state of interethnic relations according to social networks (2015). Suggested approaches and tools are not intended to identify, process and analyze scientific information data, search and explication of contexts respectively the analyzing field of knowledge.
3. The Institute of Informatics Systems named by A. P. Ershov of Siberian branch of Russian Academy of Sciences (Novosibirsk, Russia). Artificial Intelligence Laboratory: Research and development of methods and tools for analysis and visualization of heterogeneous knowledge of large information portals (2009–2011). The concept of thematic intellectual scientific Internet resource (ISIR) was suggested for informational and analytical support of the scientific and productive activity in a particular field of knowledge. ISIR technology is targeted to professionals in different fields of knowledge. It represents the development of the previous technology for building portals of scientific knowledge. The structure of the electronic

Russian-English Thesaurus on Computer Linguistics is worked out. The thesaurus is a complete and consistent system of concepts from the field of Computer Linguistics, linked by semantic relations, reflecting the position of each notion in this system.

4. The Institute for System Programming named by V. P. Ivannikov, Russian Academy of Sciences (Moscow, Russia). The Institute has developed a software system "Texterra" to use Wikipedia for enhanced search and navigation in text database. "Texterra" is an effective system from computational point of view. It uses semantic proximity measure for information search and improvement of the query results ranking. Semantic information accumulated in Wikipedia-type databases allows improving methods of automatic text processing and information retrieval. Instead of keywords, domain specific terms are used. Unfortunately, the effectiveness of such a search engine is not possible due to the lack of effective methods of indexing documents for query execution. Modern methods of calculating semantic similarity in English are well described in the following studies: [*Agirre et al.,* 2009; *Bruni et al.*, 2014; *Clark, Curran,* 2004; *Ferret,* 2010; *Mikolov et al.,* 2013; *Sahlgren*, 2006]. The tool does not support the network distributed environment.
5. University ITMO (Saint-Petersburg, Russia): Humanitariana: creation of a virtual information and resource center for the extraction of knowledge from humanitarian texts based on advanced full-text search, and functional integration of resources and services in a distributed environment (2014–2016) [*Lyapin, Kukovyakin,* 2015]. It supports operating in a network distributed environment and allows carrying out search, explication, integrated analysis of the context knowledge. More-over, it implements the possibility to build thematic trends of concept trends.

The research in the field of context knowledge is carried out by a number of specialists abroad and in Russia. Approaches, models and methods of search of information resources with use of semantic technologies are discussed in several publications [*Apanovich, Marchuk,* 2014; *Buhanovski, Vasilyev,* 2010; *Kanygin, Poltinnikova,* 2015; *Kanygin, Poltinnikova,* 2016; *Kurshev et al.*, 2002; *Taylor,* 2007; *Zagorulko, Borovikova,* 2011]. The analysis has revealed that comprehensive research, which would integrate all those approaches: from the development of search algorithms and the establishment of appropriate information systems of quantitative data accumulation to the use of those systems and algorithms in qualitative analysis of dispersed arrays of text data in heterogeneous information systems, is not being held. In addition, the research considered offers solutions that among the diversity of used approaches presuppose the existence of formalized conceptually structured base and categorical apparatus of interdisciplinary research. The last one contains thesauri, sets of interrelated keywords or domain ontology concepts, contexts, etc. relevant to the field of research.

Existing methods and their corresponding search tools, explication and analysis of contextual knowledge implemented in individual information systems and web-services do not remain well known and claimed due to the lack of meaningful information about the tool itself as well as the methods and algorithms of its use in science. Weak demand of such tools is explained due to the lack of thematic development for the selection and systematization of meta-descriptions, thesauri, subject domain ontologies that form "database and knowledge" of those tools. The other reason is implementation of a limited set of methods for presenting for such a demand, searching and interpreting information in each separate tool.

Synthetic Method of Contextual Knowledge Extraction

In this regard, for the extraction and contextual analysis of Russian full-text bases the efforts of the researchers are concentrated on the affordable distributed network environment of ITMO University. As a complex method of context data search and data analysis into this environment Synthetic method was proposed [*Kononova et al.,* 2017].

Synthetic method of contextual knowledge extraction is an integrated approach oriented to solve problems of allocation and explication of scientific content on topical directions of interdisciplinary scientific research. Synthetic method uses advanced systems of full-text and multimodal search of the network distributed environment. The method involves the extraction, expert evaluation and interpretation of contextual knowledge from large text and non-text information data. This method makes it possible to enter a new level of understanding the typologies of context knowledge, formation, objectification and updating of terminology basis of interdisciplinary research.

The developed integrated approach, named here synthetic method, displays the content analysis to a higher level. In the traditional content-analysis the target function and category analysis are of primary importance, while derived generalized text units of analysis are more important. For the synthetic method, a generalized text (with elements of multimodal information) has a priority, the resulting content — structured description of contextual knowledge is secondary. It can be said that the traditional content-analysis and proposed synthetic analysis are complementary methods and technologies of learning content and semantic information contexts. Generic text (text + multimodal information) in this case is a generator of explicated contexts and context knowledge structures. The method being developed contains:

- a primarily explication of the corpus (expert asses of the terminological base: dictionary, tutorials, manuals, program documents);
- a construction of mind maps;
- a method of representative samples selection of scientific texts of different origin and location from open sources;
- a step-by-step instruction of using combinations of well-known search methods, technologies and tools of automated extraction and study of contextual knowledge;
- an expert assessment algorithm of selected paragraph-contexts, which is original and simple for users, enables researchers to generate personal thematic collections of materials;
- construction technology of thesauri of interdisciplinary scientific direction and thematic trends within those fields;
- building of the corpus ontology and creation of thesauri of interdisciplinary scientific direction;
- a mechanism of integration of the results in selected environment, supported by recommendations on the interpretation and use.

Unlike existing developments, expected results of synthetic method applying are supposed to examine the structures and functions of knowledge in interdisciplinary research deeper as well as to use the contextual framework for interpreting scientific texts, algorithm development, software intelligent search and artificial intelligence.

The main objective of the study is to explore the explication possibilities of scientific knowledge gained from the digital information resources in a distributed network environment, using methods and technologies of extraction and analysis of contextual knowledge. To achieve this goal the following main tasks should be solved:

6. Justification of applied environment selection to ensure the objectives of the study in accordance with the established model of contextual knowledge explication;
7. Forming the semantic core (key concepts) of the subject domain of the interdisciplinary scientific direction "Digital Economy: e-governance and smart-technology";
8. Study of the possibility of using digital information resources for forming arrays of scientific texts which are relevant to subject domain (Scopus, WoS, publishing platforms, EBSCO, JSTOR, Google Academy, RINC, Kyber-Leninka, etc.); characteristics of resources, criteria influencing the choice, polithemes, etc. [*Prokudin, D., Levit, G., Hossfeld, U.,* 2017];
9. Estimation of text arrays of distributed network environment on the relevance of a given subject domain (testing on key concepts).
10. Construction of the specialized thesauri of advanced interdisciplinary research.

Several advanced directions of interdisciplinary scientific research are selected for the approbation of the synthetic method. The first stage direction is "Digital Economy: e-governance and smart technology".

Architecture and services of the network scientific environment

Distributed network environment architecture

Explication of various kinds and forms of the contextual knowledge involves the development and use of retrieval information systems with services of contextual search, different types of queries, automated extraction of knowledge from scientific texts. The authors have chosen a distributed network environment T-Libra to implement the proposed synthetic method [*Kononova et al.,* 2017]. One of its components is installed on the server of ITMO University. A distributed network environment architecture is shown below (see the Fig. 1).

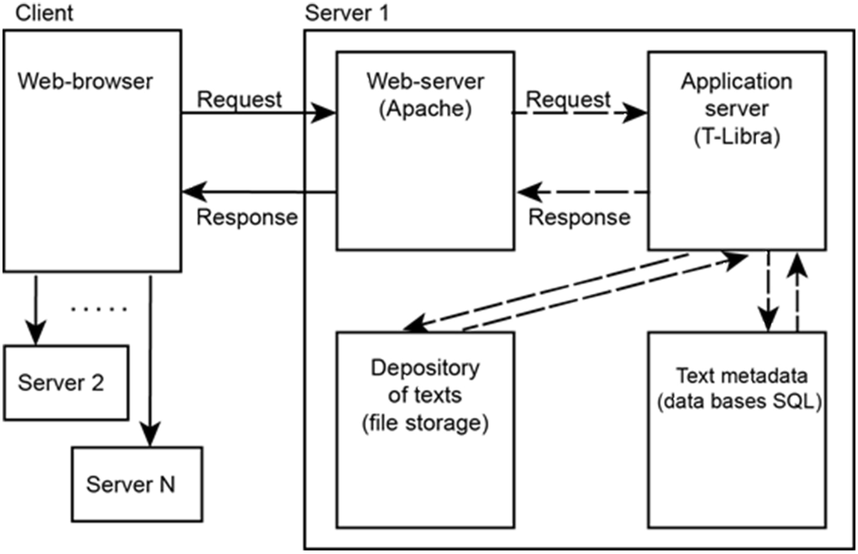

Fig. 1. T-Libra architecture

T-Libra is realized in the Internet-architecture in WWS-configuration (Web-browser / Web-server / SQL-server). It is developed in a client-server Internet/Intranet architecture: Web-browser/Web-server — Application Server/Relational DBMS, with protocols HTTP, CGI, PIPE API, ODBC. Given trends in the development of modern information space, a decentralized environment model is selected under the control of the user's browser, with a focus on Web services and Internet protocols. The web-browsers access to a plurality of independent servers. Servers are managed by different organizations. Advantage of such architecture use is in the possibility of applying the resources of several organizations. The chosen distributed network scientific environment ensured the use of tools and contextual search services; different types of queries, automated knowledge extraction from the scientific texts. The environment provides results reproducibility and results interpretability allowing to prove the relevance and practical significance of the study subject matter [*Kononova, Lyapin,* 2016]. There are three types of organizations which apply T-Libra: libraries, museums, research and educational institutions. Each organization focuses on objectives and issues corresponding to its main activity (see Table 1).

Table 1. The opportunities of T-Libra usage

Type of organization	Kind of activity	Organization
Libraries	Saving the social memory, full-text search, holding literature exhibitions, creating the full text digitization	Central State Public Library (CSPL) named by V. V. Mayakovski, Saint-Petersburg
Museums	Planning the exhibitions, expositions	Arkhangelsk local history museum, Kremlin' Museums
Institutions	Analytic research	ITMO University

Services of the network scientific environment

The multifunctional network of ITMO University scientific environment provides reproducibility and interpretability of results, allows proving the relevance and practical significance of the subject matter of the study. Scholars and research groups may work in the local network mode and in the distributed IT environment with ability to appeal to the resources of any participating organizations servers. ITMO University server provides the transparent information environment with distributed resources (informational and analytical) as well as services, free access for individual researchers, research groups, laboratories and institutions.

There are a few types of full-text search in NRE (two types of paragraph-oriented, four types of frequency-oriented searches). NRE also supports various forms of representation of query results.

Paragraph-oriented search is a search carried out on a selected set of resources based on inflection search terms (plural or singular nouns in English, single nominative nouns in Russian). Paragraph-oriented searches are represented by the varieties of work in both local and distributed environment.

Information retrieval is designed to search for and present the text up to the individual copyright paragraphs that contain user-defined terminological structure (thus explicating 'horizontal' micro-context where the key terms are the paragraph part). The author paragraph is selected as the initial unit of semantic articulation of the text.

Simple (‘single-layer’) thematic search uses one complex field for entering one or more terms (Table 2). There are combinations of terms and logical operators in any query; logical operators are join, compulsory exclusion or explicitly.

Table 2. Single-layer thematic search

Inputs. The author terminological structure	**Key term or terms connected by logical operators**
Search results	**Opportunities**
Paragraphs cluster (7 paragraphs of one document = micro context + 3 paragraphs before + 3 paragraphs after)	Viewing, on the same screen page, the corresponding resource (article, book, etc.) in file formats
A list of the relevant paragraphs	Expert evaluation of the found paragraphs
Thematic collections of micro-contexts from different documents (a list of query relevant paragraphs)	Automated assembly of thematically oriented paragraphs of the various documents in a separate file, along with bibliographic descriptions
Files	The file recording on a portable storage medium

Advanced (‘multi-layer’) *thematic search* focuses on the additional functional thematic of the request. Advanced (‘multi-layer’) thematic search is used for additional thematic focus of the query. The first layer is the base. The search field ‘layer’ (from 2 to 8 layers) is a technical tool for the isolation of a substantial aspect of interest to the user. Thus, in the query the first key-term ‘smart technology’ is specialized in connection with the ‘data’ and the ‘reliability.’ The query allows for a certain terminology field, defined by all terms of the query, to fix a semantic connection between terms related to different layers.

Frequency-based search is constructed for frequency-ranked lists of terms as well as the explication of the various “vertical” macro-context implicitly presented in a separate document, or in an array of documents. Environment supports two frequency-based searches:

— absolute frequency, which results in a frequency-ranked list of nouns included in the resource search and reduced to the initial (so-called normal) form (nominative, singular);

— relative frequency, which results in a frequency-ranked list of nouns, that includes only the paragraphs with the term predetermined by the user.

The search can be carried out simultaneously on one, two or three “baskets resources.” The result of relative-frequency query is a frequency-ranked list of nouns (‘terminogramma’). Terminogramma is a table of terms, which contains information about the absolute (in numbers) and relative (in %, ppm) frequency of terms used in papers. Implementing of contextual search services include: textual analysis of the document; description of the documents subject domain; drawing up a list of keywords; comparative analysis of subject domains, and others.

Study of the subject domain ‘Digital economy’

The preliminary stage. Terminological core of the subject domain

In order to identify clusters of contextual knowledge and to construct thematic trends of the subject domain the combination of the paragraph-oriented (‘single-layer’ query) and frequency-oriented (relative-frequency query) searches are used.

Paper considers some specific cases of the contextual knowledge explication and focuses on possibilities of the subject science fields. Experimental research base are science texts about ‘Digital Economy’ and ‘smart technology’. According to Synthetic method of contextual knowledge extraction the first experiment has objectives:

— to form a ‘term-concepts’ set of the interdisciplinary research direction ‘Digital Economy: e-governance and smart technology’;

— to identify clusters of contextual knowledge correlated with those term-concepts;

— to construct thematic trends of the subject domain on the basis of the term-concepts;

— to get automatic construction of “terminogramma” and expert analysis of the selected contexts.

The preliminary stage of the current research is the selection of the terminological core of the Digital Economy subject domain. The analysis of the National program “Digital Economy of the Russian Federation” (approved by the Decree of the Russian Federation Government dated July 28, 2017 № 1632-р) was carried out to initially identify interdisciplinary direction corpus of Digital Economy: ‘e-governance’ and ‘smart technology’. As the basic direction, the Program mentions conditions for “improving the availability and quality of public services for citizens” and among its stated objectives it highlights “creating ecosystems of digital economy of the Russian Federation, ... where effective interaction ... of the State and its citizens is ensured”. Analysis of the document allows identifying semantic core that includes the following terms related to the interdisciplinary direction as well as subject domain (Table 3). Correlating of those terms with various fields of knowledge suggests the interdisciplinary origin of ‘Digital Economy’ both as a concept in particular and the whole scientific direction in general.

Table 3. Terminological core of “Digital Economy of the Russian Federation” State Program

Term-concept (Russian — English)	**Subject domain**					
	Computer science	**Economics**	**Sociology**	**Education**	**Political science**	**Management**
Administration (Public)						+
Citizen			+		+	+
City (Smart)	+					+
Data	+		+			+
Data (Personal)			+			+
Ecosystem (of Digital Economy)	+	+	+			+
Service (State, Municipal)		+				+
Society			+			+
Storage (Center)	+					+
State					+	+
System (Information)	+					+

Analysis of the submitted semantic core reveals that Digital Economy in the field of e-Public administration continues the development of ideas and projects, which have started developing in Russia since 2005 and can be characterized by state terms (term-concepts): ‘public administration’, ‘e-governance’, ‘e-state’, ‘state information system’. In the field of both interdisciplinary directions development as ‘public administration’ and ‘Digital Economy’ the term-concept ‘public services’ (including ‘municipal services’) is more frequently used. Therefore, the directions also include such term-concepts as ‘information technology’, ‘info-communication technologies’. All term-concepts were also evolved in the context of the Strategy, approved by the Decree of the President of the Russian Federation in May 9, 2017 № 203 “About the development strategy of information society in Russian Federation at 2017–2030”.

The analysis of the corpus of digital resources

For the further study of the development of the interdisciplinary direction corpus the authors assessed the relevance of the main terms in Russian full-text digital electronic resources to the basic concepts. Those resources include: scientific publications (eLibrary, Kiber-Leninka, Socionet, East View); scientific, popular scientific and educational-methodical literature (EBS “Lan’”, East View); as well as MEDIA-newspapers and magazines (Integrum, East View). Using the built-in search tools, we get the following results (Table 4).

Table 4. Distribution of publications on major term-concepts in Russian full-text resources

Digital resources	**Term-concepts, frequency of use**			
	‘Digital Economy’	**‘Digital Economy’ & ‘State services’**	**‘State services’ & ‘Information Technology’**	**‘State services’ & ‘Digital technologies’**
eLibrary	1124	6	826	44
Cyberleninka	606	66	2942	11
Socionet	80	2	28	0
Integrum	5121	132	14398	436
EBS “Lan’”	1057	312	1085	580
East View	141	5	151	8

Analysis of the results from eLibrary and East View indicates that in both scientific and socio-political discourse, the considering term-concepts appeared in the early 2000’s, that corresponds to the beginning of the information society development in Russia as a sustainable social trend.

To evaluate the dynamics of the corpus of the reporting interdisciplinary direction search for scientific publications on the term-concepts of ‘Digital Economy’, ‘e-governance’ and ‘e-government’ was carried out in the scientific electronic library (eLibrary). The sample chosen include publications since 2005 to 2017. Dynamics of publications shows a steady growth in the use of the term-concept of ‘e-governance’ and its more frequent use in comparison with the term ‘e-Government’. The term ‘Digital Economy’ has started entering scientific discourse since 2010. The peak of its use is accounted was in 2016 and 2017.

Clarifying the terminological context of the domain core

For initial clarification of the semantic kernel of the term-concepts in the certain interdisciplinary field, a pilot study has been held where the network scientific environment installed on the server of ITMO University is used.

Information arrays are obtained from the Russian information-analytical web portal e-Library (http://elibrary.ru). The documents' set is formed by bringing the keywords 'digital economy', 'e-government', and 'e-governance' into the search bar. The advanced search is limited by the search in journals and conference proceedings as well as by the dates of publication from 2005 to 2017. More than 450 full-text documents are saved and imported into the distributed network environment of ITMO University for the complex content analysis.

Simple ('single-layer') thematic search [keyword: 'digital economy']. In total, 1690 relevant paragraphs have been found in 382 of 2973 T-Libra documents on ITMO University server. The results of paragraph-oriented search have shown, that the availability of publications related directly to the 'Digital Economy' allows carrying out an expert analysis of the paragraphs selected and forming thematic collections of scientific publications for 2005, 2011 and 2016\17. Papers for the selected period of time demonstrate changing of the domain terms' context the most accurately. This fact allows conducting the research on their basis.

Frequency-oriented search makes it possible to clarify the context of using the terms identified previously. The frequency-oriented search is carried out for the construction of terminogramma. Terminogramma is a table, that contains a column with a frequency-ranked list of words (nouns), as well as the columns with an indication of the absolute frequency (number) and the relative frequency of the term occurrence (in % ppm). The researcher is able to generate two types of frequency-oriented queries: the absolute query and the relative query. The query depth can be arbitrary. The result of the absolute query with absolute frequency (number) by the query depth 30 is represented in Table 5.

Table 5. Terminogramma: Absolute-frequency query results (fragment)

"Resources baskets"					
2005		**2011**		**2016\17**	
Word (ranking[1])	**Frequency of use**	**Word (ranking)**	**Frequency of use**	**Word (ranking)**	**Frequency of use**
System (1)	200	System (1)	516	System (1)	131
Administration (Public) (2)	111	Administration (Public) (2)	331	Administration (Public) (2)	109
Process (4)	69	Process (3)	290	*Technology (3)*	81
Project (6)	57	*Technology (5)*	185	Society (7)	72
Information (7)	56	Development (7)	143	Information (9)	66
Enterprise (8)	51	Information (9)	137	Development (11)	59
Decision (11)	47	Decision (13)	115	*Power (19)*	45
Technology (15)	42	Quality (14)	112	*Service (26)*	37

[1] The ranking is the position of a term in the terminogramma, which is defined by the frequency of the term use.

These days terminogramma demonstrates the trends: increasing interest in 'technology' as well as focus on state 'power' and 'service' in 2017.

Relative frequency-oriented query allows setting terms and their occurrence frequency in the paragraphs, where the reference term is presented. A query is done in a few stages:

1. Definition of the "reference term" and "depth of sampling frequency ranking" of the subject domain.
2. Selecting the resource array (papers) and distributing the papers into the "resource baskets" into the "resource baskets".
3. Automatic forming of a set of paragraphs, which contain user-defined reference term.
4. Constructing of the final paragraph-ranked list based on user-selected set of paragraphs or terms of the subject domain core.
5. Analysis and constructing of trends.

Relative queries have been made on five reference terms: 'Economy', 'Technology', 'State', 'Government', 'Service'. Query made on the reference term 'Service' has given no results.

Relative frequency-oriented query on the reference term 'Economy'. All three "resource baskets" of 2005, 2011, and 2016\17 have demonstrated that the main context of publications is 'administration' and 'development' issues. The focus on publications of 2005 is 'efficiency', 'process' and business 'optimization'. The focus on publications of 2011 is 'technologies' and 'regions'. The focus of publications of 2016\17 is shifted to the transformation of the state and society. It should be noted, that it is impossible to find the interaction between the term 'administration' and the terms 'power', 'government', and 'services' on the whole array of publications.

Relative frequency-oriented query on the reference term 'Technology'. In 2005 the use context of the term 'technology' was 'training' and 'business'. In 2011, the first slot was 'network', 'analysis'. In 2016\17 the term 'technology' was considered as 'service'.

Results of relative frequency-oriented queries on 'State' and 'Government' reference terms are shown in Table 6.

Table 6. Results of relative frequency-oriented queries

"Resources baskets"	Relative frequency-oriented queries	
	Reference term 'State'	Reference term 'Government'
2005	Process, Internet, Business, Technology, People	The term is missing
2011	Development, Russia, System, Necessity	Service, Administration[2] (Public), Technology, Region
2016/17	Society, Development, Technology, Transformation	Power, Administration (Public), Authority, Interaction, Population

The problem of synonyms is general. It is difficult to establish the relevance of the term context due to the difference in linguistic preferences and language peculiarities, that difference generates ambiguity and inaccuracy of the terminological base of the subject domain.

[2] The exact correspondence of the certain terms in Russian and in English is impossible, because of the ambiguity of the term 'Управление' both in Russian and in English (Administration \ Governance \ Management \ Control\ others).

The frequency-oriented search has helped to assess such content analytical indicators of the Russian scientific articles, devoted to Digital Economy, as:

— the most commonly used terms in subject domain (it is the answer to the question "What topics are there...?");

— "top-30" depth of sampling frequency ranking is the depth of the request or the occurrence of terms, in which we are interested in accounted range (it is the answer to the question "Is there anything on targeted topic?";

— connections between the most commonly used terms with the other words (it is the answer to the question "What is a semantic context of the key topics?").

Analysis of the conducted experiment results suggests that the description of the interdisciplinary direction 'The Digital Economy: 'e-governance' and 'smart technology' besides the selected term-concepts needs to be supplemented by the following ones: 'state information system', 'public service (electronic version)', and 'digital technology.'

Further studies can be pursued the objectives of expanding the resource base which they are held on, and the development of appropriate tools.

Conclusions

The conducted pilot study has proved the efficiency in the use of synthetic method of studying context knowledge on the example of the text-corpus development of interdisciplinary scientific direction 'Digital Economy: e-governance and smart technology'. The main results of the pilot study are the following ones:

— There has been revealed the possibility of the use of digital information resources for replenishment of distributed information environment in researched subject domain of the interdisciplinary scientific direction 'Digital Economy: e-governance and smart technology'. The main resources are: scientific electronic library (publications), Integrum (mass-media) and East View (scientific publications and media).

— There have been elicited the mechanisms for distributed environment to extend the semantic core (conceptual base) of the subject domain of the interdisciplinary scientific direction 'Digital Economy: e-governance and smart technology'. Conceptual base is expanded by the new term-concepts.

— It has been shown the possibility to study the dynamics of corpus of interdisciplinary scientific direction 'Digital Economy: e-governance and smart technology' using digital information resources, for example, scientific electronic library (eLibrary).

The study has demonstrated the capabilities of the toolkit and synthetic method for identifying, using and further clarifying the terminology core of the subject domain. The further development of research involves:

— supplementing the distributed information environment T-Libra with new texts;

— carrying out a full text cluster analysis to identify the contextual knowledge associated with the term-concepts of subject domain;

— forming of a semantic core connected with the direction of 'Smart technology' in the context of 'Digital Economy' development;

— revealing the dynamics of the term-concepts subject domain development of the interdisciplinary scientific direction 'Digital Economy: e-governance and smart technology' with their interactions.

References

Ageev, M.S., Dobrov, B.V., Zhuravlev, S.V., Lukashevich, N.V., Sidorov, A.V., Judina, T.N. (2002). “Tekhnologicheskiye aspekty organizatsii dostupa k raznorodnym informatsionnym resursam v universitetskoy informatsionnoy sisteme Rossii” [Technological Aspects of Access Organization to Heterogeneous Information Resources in the University System of Russia], *Russian Digital Libraries Journal*, no. 5(2), pp. 1–13 (in Russian).

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches, in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL ‘09),* Association for Computational Linguistics, Stroudsburg: PA, USA, pp. 19–27.

Agosti, M., Benfante, L., Orio, N. (2003). “IPSA: A Digital Archive of Herbals to Support Scientific Research”, in: T.M.T. Sembok, H. B. Zaman, H. Chen, S. R. Urs, S. H. Myaeng (eds), *Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. ICADL. Lecture Notes in Computer Science*, Berlin; Heidelberg: Springer, 2911, pp. 253–264, DOI:10.1007/978-3-540-24594-0_24

Apanovich, Z.N., Marchuk, A.M. (2014). “Podhody k normalizacii slovarey i ustanovleniyu identichnosti sushhnostey pri obogashhenii kontenta nauchnyh baz znaniy” [New Approaches Towards Normalization of Dictionaries and Establishment of the Identity of Entities in Content Enrichment of Scientific Knowledge Bases], in: *XIV nacional’naya konferentsiya po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2014 (24–27 oktjabrya 2014 g., Kazan’, Rossiya): Trudy konferentsii.* [Proceedings of the XIV National Conference on Artificial Intelligence with International Participation, 24–27 October, 2014], v 3 tomakh, Kazan’: Izd-vo RIC «Shkola», vol. 1, pp. 92–100 (in Russian).

Borgman, C.L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press.

Borgman, C.L. (2009). “The Digital Future Is Now: A Call to Action for the Humanities”, *Digital Humanities Quarterly*, vol. 3, no. 4. Available at: http://digitalhumanities.org/dhq/ (date accessed: 21.02.2018).

Borodkin, L.I. (2008). “Prioritety sovremennoy istoricheskoy informatiki: tehnologii e-Science” [Priorities of Modern Historical Information: Technologies of e-Science], in: *Krug idey: mezhdistsiplinarnyye podkhody v istoricheskoy informatike. Trudy X konferentsii Assotsiatsii “Istoriya i komp’yuter”* [The Circle of Ideas: Interdisciplinary Approaches in Historical Informatics], Moskva, pp. 5–15 (in Russian).

Bruni, E., Tran, N.K., Baroni, M. (2014). “Multimodal Distributional Semantics”, *Journal of Artificial Intelligence Research,* vol. 49, pp. 1–47, DOI:10.1613/jair.4135

Buhanovski, A.V., Vasilyev, V.N. (2010). “Sovremennyye programmnyye kompleksy komp’’yuternogo modelirovaniya e-science” [Modern Software Complexes of Computer Modelling of e-Science], *Izvestija vysshikh uchebnykh zavedeniy. Priborostroyeniye* [Proceedings of Higher Schools. Instrument-making Industry], vol. 53, no. 3, pp. 60–64 (in Russian).

Burda, D., Teuteberg, F. (2013). “Sustaining Accessibility of Information through Digital Preservation: A Literature Review”, *Journal of Information Science*, vol. 39, no. 4, pp. 442–458. DOI:10.1177/0165551513480107

Chernij, A.V., Tuzovskij, A.F. (2009). “Razvitiye informatsionnoy sistemy organizatsii s ispol’zovaniyem semanticheskikh tekhnologiy” [The Development of Organization Information System with Semantic Technologies Use], in: *Znaniya — Ontologii — Teorii, Novosibirsk, 20–22 oktyabrya 2009* [Knowledge — Ontology — Theory. Proceedings of the Conference, Novosibirsk, October 20–22, 2009], Novosibirsk: ZAO “RIC Prajs-Kur’er”, vol. 2, pp. 52–59 [in Russian].

Clark, S., Curran, J.R. (2004). “The Importance of Supercharging for Wide-coverage CCG Parsing”, in: *Proceedings of the 20th international conference on Computational Linguistics (COLING ‘04).* Association for Computational Linguistics, Stroudsburg, PA, USA DOI: https://doi.org/10.3115/1220355.1220396

Digital Libraries and Information Access: Research Perspectives (2012), G. G. Chowdhury, Sch. Foo (eds.), Chicago, Illinois: Neal-Schuman/ALA.

Domingue, J., Fensel, D., Hendler, J.A. (2011). *Handbook of Semantic Web Technologies*, Heidelberg; Dordrecht; London; N.Y.: Springer.

Ferret, O. (2010). "Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus", in: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Languages Resources Association (ELRA)*, pp. 3338–3343. Available at: https://pdfs.semanticscholar.org/4501/9f898aa59fba708761f1c7f2854c8aa45138.pdf, last accessed 2018/02/21 (date accessed: 16.05.2019).

Kanygin, G.V., Poltinnikova, M.S. (2016). "Kontekstno-orientirovannyye ontologicheskiye metody v sotsiologii" [Context-oriented Ontological Methods in Sociology], in: *Trudy SPIIRAN* [Proceedings of SPIIRAS], iss. 48, pp. 107–124 (in Russian).

Kanygin, G.V., Poltinnikova, M.S. (2015). "Social'noye znaniye i voprosy razrabotki ego instrumentov" [Social Knowledge and the Development of its Tools], in: *Peterburgskaya sotsiologiya segodnya. Sbornik trudov Sotsiologicheskogo instituta RAN,* [Petersburg Sociology Nowdays. Proceedings of the Sociological Institute of RAS], S.-Peterburg, pp. 359–373 (in Russian).

Kononova, O., Lyapin, S. (2016). "Using the Contextual Search for the Organization Scientific Research Activities", in: *Communications in Computer and Information Science*, iss. 674, pp. 392–399. DOI: 10.1007/978-3-319-49700-6_38

Kononova, O., Prokudin, D., Liapin. S. (2017). "Sinteticheskiy metod izvlecheniya kontekstnogo znaniya v russkoyazychnoy sotsial'no-gumanitarnoy sfere: kompleksnyy podhod" [The Synthetic Method of Contextual Knowledge Explication in the Russian Socio-Humanitarian Sphere: An Integrated Approach], in: *Materialy XX Ob'yedinennoy konferentsii "Internet i sovremennoye obshchestvo" (IMS-2017), Sankt-Peterburg, 21–23 Iyunya 2017* [Proceedings of the XX joint conference "Internet and Modern Society" (IMS-2017), Saint-Petersburg, June 21–23, 2017], S.-Peterburg, pp. 53–67. Available at: http://ojs.ifmo.ru/index.php/IMS/article/view/503 (date accessed: 21.02.2018) (in Russian).

Kurshev, E.P., Osipov, G.S., Ryabkov, O.V., Sambu, E.I, Solovjeva, N.V., Trofimov, I.V. (2002). "Intellektual'naya metapoiskovaya sistema"[Intelligent Metasearch System], in: *Trudy mezhdunarodnogo seminara „Dialog'2002. Komp'yuternaya lingvistika i intellektual'nyye tehnologii"* [Proceedings of the International Workshop 'Dialog-2002'], Moskva: Nauka, pp. 320–330 (in Russian).

Laakso, M. (2014). "Green Open Access Policies of Scholarly Journal Publishers: a Study of What, When, and Where Self-archiving Is Allowed", *Scientometrics*, vol. 99, no. 2, pp. 475–494. DOI:10.1007/s11192-013-1205-3

Lyapin, S. Kh., Kukovyakin, A.V. (2015). "Tematicheskiye kollektsii polnotekstovykh zaprosov dlya izucheniya kontekstnogo znaniya v gumanitarnoy sfere" (proekt Humanitariana) [Thematic collections of full-text queries for learning contextual knowledge (project Humanitariana)], in: *Materialy XVIII Ob'yedinennoy konferentsii "Internet i sovremennoye obshchestvo" (IMS-2015),* [Proceedings of the XVIIIth Jont Scientific Conference 'Internet and Modern Society' (IMS-2015)], S.-Peterburg: ITMO University, pp. 216–224 (in Russian).

Lyapin, S. Kh., Kukovyakin, A.V., Kudryavtseva, M.V. (2016). "Ispol'zovaniye instrumentov elektronnoy biblioteki dlya vyyavleniya ponyatiyno-tematicheskikh trendov" [The Use of e-Library for the Identification of Conceptual and Thematic Trends], in: *Materialy XIX Ob'yedinennoy konferentsii "Internet i sovremennoye obshchestvo" (IMS-2016)* [Proceedings of the XIXth Joint Conference 'Internet and Modern Society' (IMS-2016)], S.-Peterburg: ITMO University, pp. 70–86 (in Russian).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). "Distributed Representations of Words and Phrases and Their Compositionality", in: *Proceedings Advances in Neural Information Processing Systems*, iss. 26, pp. 3111–3119.

Nielsen, H.J., Hjørland, B. (2014). "Curating Research Data: the Potential Roles of Libraries and Information Professionals", in: *Journal of Documentation*, vol. 70, no. 2, pp. 221–240, DOI:10.1108/JD-03-2013-0034

Osipov, G.S., Tihomirov, I.A., Smirnov, I.V. (2004). "Intellektual'nyy poisk v global'nykh i lokal'nykh vychislitel'nykh setyakh i bazakh dannykh" [Intelligent Search in Global and Local Computational Networks and Databases], in: *Programmnye sistemy: teoriya i prilozheniya. Trudy mezhdunarodnoy konferentsii* [Program Systems: Theory and Application. Proceedings of the International Conference], IPS RAN, Pereslavl'-Zalesskiy, Moskva: Fizmatlit, vol. 1, pp. 21–23 (in Russian).

Prokudin, D., Levit, G., Hossfeld, U. (2017). "Selection Methods of Digital Information Resources for Scientific Heritage Studies: A Case Study of Georgy F. Gause", in: *Internet and Modern Society: Proceedings of the International Conference IMS-2017* (St. Petersburg; Russian Federation, June 21–24, 2017), ACM International Conference Proceeding Series, N.Y.: ACM Press, pp. 69–74. DOI: 10.1145/3143699.3143739

Sahlgren, M. (2006). *The Word-space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Space*, PhD Thesis, Univ., Stockholm.

Saifa, H., Heb, Y., Fernandeza, M., Alani, H. (2016). "Contextual Semantics for Sentiment Analysis of Twitter", in: *Information Processing & Management*, vol. 52, no. 1, pp. 5–19. DOI: 10.1016/j.ipm.2015.01.005

Scanlon, E. (2014). "Scholarship in the Digital Age: Open Educational Resources, Publication and Public Engagement", *British Journal of Education Technologies*, vol. 45, pp. 12–23. DOI:10.1111/bjet.12010

Tao, F., Hou, K.L., Han, J., Zhai, C., Cheng, X., Danilevsky, M., Desai, N., Ding, B., Ge Ge, J., Ji, H., Kanade, R., Kao, A., Li, Q., Li, Y., Lin, C., Liu, J., Oza, N., Srivastava, A., Tjoelker, R., Wang, C., Zhang, D., Zhao, B. (2013). "EventCube: Multi-dimensional Search and Mining of Structured and Text Data", in: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'13),* New York: ACM, USA, pp. 1494–1497. DOI: http://dx.doi.org/10.1145/2487575.2487718

Taylor, W.P. (2007). *A Comparative Study on Ontology Generation and Text Clustering Using VSM, LSI, and Document Ontology Models*. Clemson University.

Turney, P.D., Pantel, P., et al. (2010). "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188. DOI:10.1613/jair.2934

Weller, M. (2011). *The Digital Scholar: How Technology Is Transforming Scholarly Practice*, London: Bloomsbury Academic, DOI:10.5040/9781849666275

Zagorulko, Y.A., Borovikova, O.I. (2011). "Postroyeniye mnogoyazychnogo tezaurusa predmetnoy oblasti sredstvami tekhnologii sozdaniya portalov nauchnykh znaniy" [Constructing Multilingual Thesaurus of Subject domain by Technology of the Development of Scientific Knowledge Portals], in: *Vserossiyskaya konferenciya s mezhdunarodnym uchastiyem 'Znaniya — Ontologii — Teorii' (ZONT-2011)* [All-Russian Conference with International Participance 'Knowledge — Ontology — Theories], Novosibirsk. Available at: http://elib.ict.nsc.ru/jspui/bitstream/ICT/1380/1/3OHT 7.pdf (date accessed: 21.02.2018) (in Russian).

Zhizhimov, O.L., Molorodov, Y.I., Pestunov, I.A., Smirnov, V.V., Fedotov, A.M. (2011). "Integratsiya raznorodnykh dannykh v zadachakh issledovaniya prirodnykh ekosistem" [Heterogenous data integration for nature ecosystems research], in: *Vestnik NSU. Series: Information Technologies* [Bulletin NSU. Series: Information Technologies], vol. 9, no. 1, pp. 67–74. Available at: http://www.nsu.ru/xmlui/handle/nsu/323, (date accessed: 21.02.2018) (in Russian).

Извлечение контекстных знаний: терминологический ландшафт цифровой экономики

Ольга Витальевна Кононова

Кандидат экономических наук,
доцент Национального исследовательского университета ИТМО
e-mail: kononolg@yandex.ru

Дмитрий Евгеньевич Прокудин

Доктор философских наук, доцент Санкт-Петербургского государственного университета,
аналитик Национального исследовательского университета ИТМО
e-mail: hogben.young@gmail.com

Елена Евграфовна Елькина

Кандидат философских наук,
доцент Национального исследовательского университета ИТМО
e-mail: e.e.e.1@mail.ru

Статья основана на докладе, сделанном авторами на «Международной конференции по материалам, прикладной физике и инженерии» (ICMAE-2018), проходившей в штате Мадхья-Прадеш, Индия, 3–4 июня 2018 г. Цель исследования состоит в демонстрации эффективности применения синтетического метода для извлечения контекстных знаний из распределенной информационной сетевой среды и их анализа в целях изучения новых тенденций в развитии научных направлений. Основная исследовательская задача заключется в выявлении тенденций в формировании корпуса «Цифровая экономика» как развивающегося междисциплинарного научного направления. В настоящей статье авторы предлагают использовать для извлечения контекстных знаний синтетический метод, который сочетает в себе различные подходы и инстсрументы цифровых гуманитарных наук. Этот метод позволяет осуществить отбор корпуса релевантных текстов из распределенной цифровой среды, экспликацию контекстных знаний и определить тезаурус нового междисциплинарного научного направления «Цифровая экономика: электронное правительство и умные технологии». Материалы исследования подтверждают возможность применения синтетического метода для изучения тенденций в развитии тезаурусов междисциплинарных научных областей. Результаты исследования могут быть использованы для разработки междисциплинарной онтологии «Цифровая экономика».

Ключевые слова: цифровая экономика, электронное правительство, контекстные знания, контекст, синтетический метод, «умные» технологии, распределенная цифровая среда, информационные ресурсы, междициплинарные научные направления.

Благодарность

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках научного проекта № 18-011-00923 («Разработка комплексного подхода к анализу развития терминологической базы развивающихся междисциплинарных направлений научных исследований в распределенной сетевой среде»).