The Approach to the Meta-description of the Interdisciplinary Research Terminological Landscape

O.V. Kononova[1][0000-0001-6293-7243] and D.E. Prokudin [1,2][0000-0002-9464-8371]

[1] St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University), 199034, St. Petersburg, Russia
[2] Saint Petersburg State University, Saint Petersburg, 7/11 Universitetskaya Emb., 199034, Russia
kononolg@yandex.ru, hogben.young@gmail.com

Abstract. The paper is aimed at conducting comprehensive research on methods and tools for searching and explicating contexts of arrays of scientific information, visualizing hierarchical and associative relations between terms considering the context. It presents the findings indicated that the both, a terminological landscape and a meta-description construction, are necessary and useful for most of the interdisciplinary research, for further contextual knowledge explication but also forecast of the societal impact of studies. Research has focused on the ontological approach to the structured description of contextual knowledge and the scientific interdisciplinary domain thesaurus. There is proposed an ontological approach to the structured description of contextual knowledge, structure, and meta-description of the thesaurus of individual interdisciplinary are-as as part of the study on the development of an integrated approach to the analysis of the terminological base of developing interdisciplinary research. The structure of the thesaurus and meta-description of its elements are proposed to be formed on the basis of the Dublin Core Metadata elements (Dublin Core Metadata Element Set). It allows using the thesaurus for automated search and identification of contextual knowledge by search engines. In the applied aspect the paper considers the method of creation of open-access electronic archives for the purpose of further replenishment, systematization, and study of contextual knowledge. Description of contextual knowledge based on Dublin Core enables to automate the exchange of meta-descriptions by using the standard OAI-PMH protocol.

Keywords: Contextual Knowledge, Synthetic Method, Thesauri, Meta-Description, Dublin Core, OAI-PMH.

1   Introduction

The development of interdisciplinary approaches is increasingly necessary for most of disciplines, both for further knowledge discovery and societal impact of discoveries [19]. There are many studies in the research of interdisciplinarity. Some of them, such as Keisuke Okamura, say that interdisciplinarity is statistically significantly and positively associated with research impact by focusing on highly cited paper clusters known as the research fronts (RFs) [18] or as Carusi C. and Bianchi G. suggest quantifying interdisciplinarity of journals by exploiting relation between scholars and journals where such scholars do publish [20]. The others such as, for instance, Raimbault J. [19] write that patterns of interdisciplinarity in science can be quantified through diverse complementary dimensions on the base of a methodology combining citation network analysis and semantic analysis or, for instance, as Giovanni Abramo, Ciriaco Andrea D'Angelo, Lin Zhang [21] use at the same time two bibliometric approaches to the measurement of interdisciplinary research: the analysis of disciplinary diversity in the reference list of publications; the disciplinary diversity of authors of publications. To analyze trends and contexts of scientific studies in sphere of gamification the researchers Jonna Koivisto, Juho Hamari [22] have used exclusively the expert research method without the application of automation of the analyze process itself or any information systems.

The aim of the study conducted by the authors is the development of science-research technologies for extracting, processing and analyzing contextual knowledge from information sources of various nature ensuring the effectiveness of scientific research. One of the main tasks is to develop specialized thesauruses of the latest interdisciplinary scientific research in the digital economy area: e-government and smart technologies; socio-ethical aspects of the digital economy – gamification in the social, scientific and educational fields; informatization of scientific activity.

As the main approach to the study for the terminological base formation, the authors' technique (Synthetic Method) of automated extraction and explication of contextual knowledge using a hybrid query to analyze and explicate contexts from information resources of text modality [4, 15] is used. The method involves the extraction, expert evaluation, and interpretation of contextual knowledge from a large massive of text resources. Expected results of the 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 Synthetic Method uses the combination of cascading search (the results of one query are automatically included in the search image of another query) and thesaurus search (automatically expansion of the cultural contexts during the query), the combinations of multi-layer thematic paragraph-oriented (with a layers' variation) and frequency-oriented queries. The method is aimed at clustering and statistical processing of scientific text, allows creating collections of thematically relevant contexts, correlating the obtained contexts with knowledge areas, as well as identifying contexts of the use of some term-concepts and keywords.

Contextual knowledge in the study is considered as a form of interpretation of texts and their contexts obtained because of search queries and explication of their meanings with contextual search technologies and tools. Information resources from scientific sources, as well as socio-political discourse publications reflected the main trends in social development are used.

Using the Synthetic Method, the following steps have been sequentially realized:

— the selection of digital information resources for further text extraction and contexts' explication (scientific and socio-political discourses);
— the identification of the basic term-concepts of the studied subject domains;
— the added search based on the identified term-concepts in the texts from selected digital information resources;
— the formations of the text corpuses, which are relevant to the research topics with the expert evaluation method;
— the construction and description of the thesauruses of the interdisciplinary scientific domains;
— the integration of the results supported with recommendations on the interpretation and use.

In order to evaluate the dynamics of research interest to the Digital Economy subject domain in such aspects as e-Governance, Smart City, Gamification Technology, the study proposes a review of the corpus of Russian and foreign scientific publications, as well as media publications. Analysis of the terminological landscape of scientific publications and media allows the following:

— to highlight promising proposals and practices on gamification using in the development of the city information space;
— to form a terminological base and thesaurus of the interdisciplinary scientific direction "Social and ethical aspects of the Digital Economy: gamification in the social, scientific and educational subject domains";
— to compare Russian and Global trends of gamification using in the development of the modern urban information environment;
— to focus the attention of researchers on a number of scientifically significant results;
— to identify the development trajectories of promising scientific areas.

At this stage of the research, the thesauruses of the terminological base (as one of the forms of presenting contextual knowledge and as a formal set of the identified contexts' descriptions) of the interdisciplinary scientific areas "The Digital Economy: Electronic Government and Smart Technologies" and "Socio-ethical aspects of the digital economy: gamification in the social, scientific and educational fields" are ready. The task of creating a structured description of the contextual knowledge, structure and meta-description of the domain terminology base has required consideration of various approaches to the thesaurus description.

2 Approaches to a Structured Description of Contextual Knowledge

The process of sequential explication of contextual knowledge to identify the terminological base of the studied interdisciplinary scientific areas has led to the refinement of the concept of contextual knowledge, as well as the creation of the typology of contexts of text modality. The choice of the ontological approach for solving the problem of a structured description of contextual knowledge, a structure and a meta-description of the terminological base thesaurus of interdisciplinary scientific domains was the logical result of the analysis of modern concepts and approaches to the presentation of ontologies and their use in practice.

There are various approaches for defining both the term ontology itself and the classification of ontologies. According to Y. Lipuntsov [6], depending on the goals and the nature of the applications in which they are used, various strategies for structuring knowledge and presenting information can be applied. One of the most popular solutions for representing thesaurus as ontology is the WordNet project [10, 16]. It is used to build large-scale thesauruses, for instance, the Russian language as a whole (RussNet project, http://project.phil.spbu.ru/RussNet). Moreover, similar resources are often used in order to create individual dictionaries, lists of key words and expressions for the solving scientific particular problems [7]. To view and edit WordNet-thesauruses saved in XML format, the freeware VisDic program (http://nlp.fi.muni.cz/projekty/visdic/) has been developed.

The format of the thesaurus description depends on the standard used. Each of the existed standards offers its own way of representing the thesaurus, but always as finite lists of the objects of several determined types and relations among them (ISO 27881986, ISO 27881986, ANSI/NISO Z39.191993, GOST 7.25-2001, GOST 7.24-2007). Some standards regulate the presentation format of the thesaurus in both machine-readable and textual forms (for example, ANSI/NISO Z39.191993). For example, GOST 7.25-2001 establishes the rules for its development, structure, composition and form of presentation", for instance, establishes the development rules, structure, composition and presentation form of information thesauruses focused on the use of Russian vocabulary and developed in the framework of automated information systems and networks of scientific and technical information [2].

There are various approaches to describe the thesaurus in relation to its use in information systems. A sufficiently detailed review of such approaches was made by M. Kh. Nguyen and A. S. Adzhiev [8]. All approaches considered by these authors are based on the final implementation of the thesaurus as a linguistic system. In the framework of our study, linguistic approaches and methods to the thesaurus formation are not directly used.

The authors consider the ontology as a system consisting of term-concepts (without their definitions), their attributes, and the relations which are necessary, according to Y. Lipuntsov [6], to limit the interpretation and use of term-concepts. The scientific thesaurus is presented in the study as a list of term concepts selected from a corpus of text documents that are relevant to the interdisciplinary research domains. The list of term-concepts is able to reflect the conceptual aspects of the subject area through a

structured description of the elements. Thereby, the thesaurus of the interdisciplinary research domains is an ontology with the thesaurus semantic that defines the relationship between term-concepts (synonymy, hierarchy, association) and the attributes, which are characteristic for conceptual schemes. The composition of the attributes and the relationships between the term-concepts are not fixed; they are determined by developers and depend on the construction purpose. Such an interpretation of the thesaurus of the interdisciplinary research domains allows representing the internal organization of each separate topic of every domain and use the thesaurus in the solving tasks such as scientific search and so as explication of contexts. For a formal presentation of the terminological base of an interdisciplinary scientific research domain, it is sufficient to use simple relations between the basic terms (key term-concepts) and related terms (relative term-concepts) identified through the explication of contextual knowledge, as well as the synonymy relations. In addition, the structure of the thesaurus description takes into account, besides the main subject area, the presence of the other areas (subdomains) involved in the formation of interdisciplinary scientific research domains.

3   Thesaurus Meta Description

An ontological approach was used to describe the structure of the domain thesaurus, which, in addition to defining relationships among term-concepts, involves the semantic expanding that defined relations of synonymy, hierarchical, and association. In the framework of the approach, when ontology is interpreted as a system consisting of plurality of concepts (classes and subclasses defined by key supportive and relative term-concepts), their attributes and relations (connections), the authors describe the thesaurus of the interdisciplinary scientific research domains using the list of key term-concepts derived from scientific papers and media documents in accordance with advanced search queries which include a list of subject areas. Descriptions of key term-concepts and contexts of their use define the content of scientific discourses of the subject domain. This approach allows determining the relation between subject domains, key term-concepts, as well as related term-concepts. At the same time, the application of the synthetic method, which involves the phased explication of contextual knowledge from the relevant arrays of information resources and bases on the results of frequency-oriented queries, has been allowed ranking the significance of the term-concepts even at the stage of thesaurus formation.

To describe the structure of the thesaurus and its terms, a set of 14 Dublin Core Metadata Element Set (DCMES) metadata elements was taken [3]. The combination of values of these elements was used as a structured description of thesaurus elements that correlated with the general approach of using this specification to text information resources [1, 11–14, 17]. This approach enables us to present the thesaurus in a form suitable for computer representation and for automated search and identification by search engines.

In the proposed approach, the thesaurus element is a key term-concept and has thesaurus type of context. The meta-description of each term-concept consists of the following Dublin Core metadata set:

dc.title – name of the term-concept;

dc.subject – subject area(s) (for instance, according to the classifier GRNTI);

dc.description – description of the term concept sources;

dc.contributor – author(s) (used if available);

dc.date – a date when the term concept was included in the thesaurus;

dc. modified – date of the last modification (modification) of the thesaurus element;

dc.type – type of term-concept or connection (key term-concept, relative term-concept, synonym, others);

dc.format – term-concept representation format;

dc.identifier – identifier of the term-concept (for instance, a link on the term-concept description on the network resource);

dc.source – name of the thesaurus;

dc.coverage – to establish the relationship of association, hierarchy or synonymy between the described key term-concept and relative term-concepts (a set of elementary pairs "key term-concept – relative term-concept"; the number of pairs is unlimited). If "synonym" is indicated in the dc.type field, then this field defines the synonymy relationship in a pair of term-concepts;

dc.relation – related elements, which, for instance, may include links to files or links to third-party resources.

In the framework of the study, the elementary pair "key term-concept – relative term-concept" is a thematic context built by analytical IS. One thesaurus record contains one or more of these pairs generated as a result of a number of relative frequency queries and in accordance with the obtained frequency values.

Based on the technique of applying the automated extraction and explication of contextual knowledge using a hybrid query, the thesaurus of the categorical and terminological base of the interdisciplinary scientific domain "Digital Economics: e-government and smart technologies" included 50 term-concepts. There were four semantic groups identified in the study: "Power", "Subjects of Relations", "Functional Areas", "Technologies." For their structured description, the proposed approach has been applied. The results are reflected in the following examples:

Example 1. The term-concept "e-government"
dc.type: key term-concept
dc.title: e-government
dc.subject: public administration
dc.description: the term is obtained as a result of a cascade query on an array of relevant texts of articles from Russian scientific journals for the requests "digital economy" (2016, 2017) and "digital technologies" (2011, 2017, 2018, eLibrary); corps of texts from the Federal and Regional press at the request of the digital economy (2017, Integrum).
dc.date: 11/05/2018

dc.source: digital economy: e-government and smart technologies
dc.language: russian
dc.type: relative term-concept
dc.coverage: digital economy
dc.coverage: digital technology
dc.coverage: information technology
dc.coverage: public administration
dc.coverage: smart technology

Example 2. The term-concept "digital technology"
dc.type: key term-concept
dc.title: digital technology
dc.subject: public administration
dc.description: the term is derived from a relative frequency query on an array of relevant contexts formed from body of texts of articles from Russian-language scientific journals for the requests "digital economy" (2016, 2017) and "digital technologies" (2011, 2017, 2018 gg) obtained from the Scientific Electronic Library; cases of texts from the federal and regional press on the request "digital economy" (2017), received from the Integrum electronic system
dc.date: 11/05/2018
dc.source: digital economy: e-government and smart technologies
dc.language: russian
dc.type: relative term-concept
dc.type: synonym
dc.coverage: information technology

4 Systems for Presentation of a Structured Description of Thesaurus Elements

In the applied aspect, taking into account the capabilities of the synthetic method, it has been proposed to create electronic archives of open access for the accumulation and systematization of contextual knowledge. Electronic archives can serve for the further replenishment, systematization, and study of contextual knowledge both within the framework of this project and for research purposes of the entire scientific community. Open access to the description of the thesauruses, a type of contextual knowledge based on the Dublin Core standard, provides:

— computer representation of contextual knowledge in information systems;
— ability to translate contextual knowledge into other presentation formats (for instance, XML with subsequent export to formats of other systems [9].)

To solve the problem of computer representation of the thesaurus, various software platforms implemented with the Dublin Core standard have been considered. First of all, those kinds of software enable to create electronic archives of open access. They include such popular solutions as DSpace (https://duraspace.org/dspace/), Eprints

(https://www.eprints.org/uk/), Fedora (https://duraspace.org/fedora/), VIVO (https://duraspace.org/vivo/), INVENIO (https://invenio-software.org) and some others (http://www.openarchives.org/pmh/tools/). The most popular ones are the DSpace and Eprints platforms, on which the vast majority of institutional repositories operate (including Russia) [5]. In addition to the standard implementing for metadata presentation with a set of Dublin Core elements, these solutions have the ability to automatically exchange metadata using the standard OAI-PMH protocol that allows gathering the metadata collections from installations of various systems, their further accumulation (aggregation), and applying in search tasks.

However, the installation and configuration of these systems require working with the operating system, and it is necessary for this to install additional software. Therefore, with experimental purposes, the open-source software Open Journal Systems (OJS, https://pkp.sfu.ca/ojs/), which is a complete publishing cycle platform for publishing electronic journals, has been chosen as a palliative solution. This system is easier to install and configure, and it is able to work on most shared hosting. Moreover, OJS supports the Dublin Core metadata presentation format and acts as an OAI-PMH provider. With experimental purposes, descriptions of several term-concepts obtained during the study were introduced into the OJS installation. To verify the correct display of metadata, we have used the installation of the Open Harvester Systems (OHS, https://pkp.sfu.ca/ohs/), which acts as a server using the OAI-PMH protocol. OHS is configured to aggregate metadata from any information resources that play the role of providers. An example of the presentation of metadata is shown in Figure 1.

Fig. 1. Description of the Dublin Core term obtained from the OHS aggregator

5   Computer-Readable Thesaurus Presentation Model

The proposed approach to a structured meta-description of the thesaurus lets us use various information systems supported the Dublin Core metadata format and the OAI-PMH automated metadata exchange protocol. This makes it possible to create and maintain a platform-independent research network, in which any research group can create its own thesaurus either in the existing system or use some own installation. At the same time, with research purposes, it can be possible to use both existing and own aggregators (collectors) of metadata from various systems in which thesauruses or sets of terms are presented (Fig. 2).
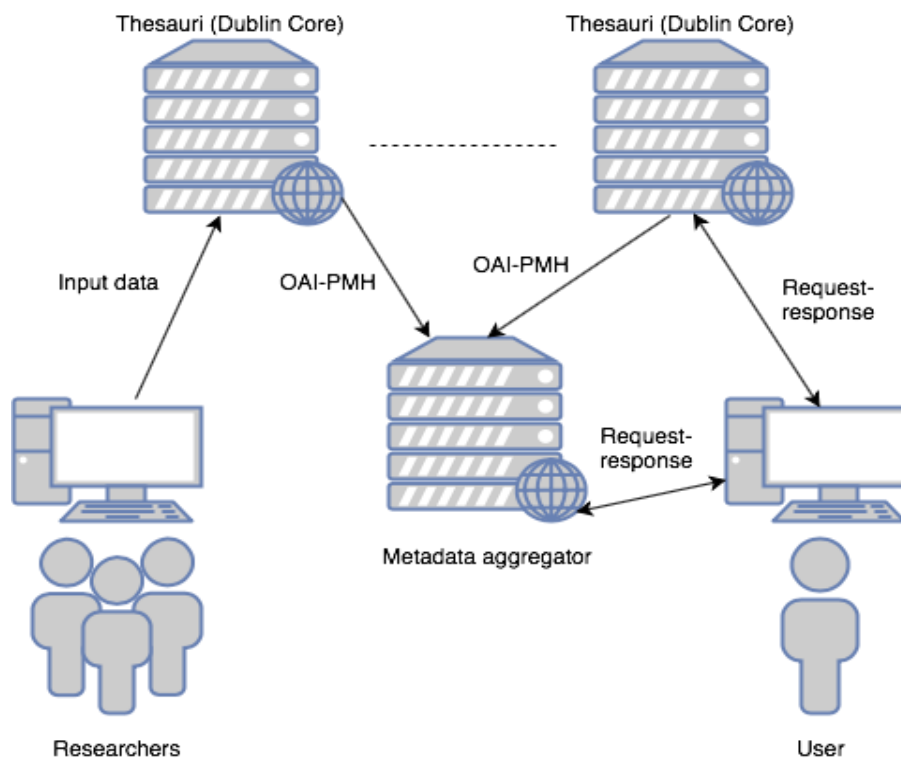
Fig. 2. Model of collective interaction in a distributed environment

Using the metadata aggregator allows not only to consolidate the descriptions of terms from distributed thesauruses, but also the metadata of publications which were presented in various electronic network resources (repositories, electronic journals, electronic libraries, etc.). It makes it possible

— present in one information space both a terminological base and descriptions of texts;
— search texts which are relevant to certain term-concepts;
— clarify the conceptual apparatus of scientific areas by analyzing the metadata of the selected texts directly.

At the same time, thematic metadata collections can install in the aggregator settings, thereby limiting the subject of aggregated texts and increasing the relevance of arrays of aggregated texts to the studied thesaurus. The approach was implemented during the construction of the thesauruses "The Digital Economy: Electronic Government and Smart Technologies" and "Socio-ethical aspects of the digital economy: gamification in the social, scientific and educational fields" based on OJS (http://ojs.iculture.spb.ru) and an aggregator based on OHS

(http://ohs.iculture.spb.ru). In addition to the thesaurus, metadata from 30 Russian-language open repositories were aggregated.

6   Conclusions

The proposed methodology for studying contextual knowledge makes it possible to sequentially extract contexts and thereby identify and form the terminological base of interdisciplinary scientific domains. The terminological base of interdisciplinary scientific domains will be replenished with new contexts and refined after processing the results of the further hybrid requests. The proposed Dublin Core-based contextual knowledge description format will provide:

— the construction of distributed replenished formal thesauruses using standard freely distributed software;
— the use of thesauruses and its elements for automated search and identification of contextual knowledge by search engines;
— archiving and open access to contextual knowledge;
— the possibility of automated exchange using the OAI-PMH protocol for aggregation of meta-descriptions of thesauruses in systems acting as OAI-PMH servers.

Further research suggests:

— to build formal thesauruses based on the Dublin Core specification, reflecting the results of the study;
— to provide open access to the thesauruses of interdisciplinary research areas;
— to systematize and classify software for processing contextual knowledge, develop their presentation and description based on the Dublin Core specification, and compile an accessible machine-readable catalog of these systems.

7   Acknowledgement

References

1. Volohin, O.M.: The Technological Model of the Library Information Portal Creation based on Dublin Core Metadata. In: D.Chizhevsky Kirovograd Regional Universal Research Library. Kirovograd, Ukraine (2003), http://www.gpntb.ru/win/inter-events/crimea2003/trud/tom1/sec/Doc64.HTML, last accessed 2019/11/05.

2. GOST 7.25-2001. Tezaurus informacionno-poiskovyj odnojazychnyj. Pravila razrabotki, struktura, sostav i forma predstavlenija. Vzamen GOST 7.25- 80; Vved. 2002-07-01. IPK Izd-vo standartov, Moscow (2001).

3. GOST R ISO 15836-2011 Informacija i dokumentacija. Nabor jelementov metadannyh Dublin Core. Nacional'nyj standart Rossijskoj Federacii. Moskva, Standartinform (2014).

4. Kononova, O.V., Lyapin, S.Kh., Prokudin, D.E.: Studying the Interdisciplinary Terminological Landscape of Digital Economy with the Use of Contextual Analysis Tools. International Journal of Open Information Technologies 6(12), pp. 57–66 (2018). http://injoit.org/index.php/j1/article/view/648/632, last accessed 2019/11/05.

5. Kudim, K.A., Proskudina, G.Yu, Reznichenko, V.A.: Comparison of repository systems EPrints 3.0 and DSpace 1.4.1. In: Trudy 9 vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii» - RCDL2007. Perejaslavl'-Zalesskij, 15–18 oktjabrja 2007 goda (2008), https://dspace.nsu.ru/xmlui/bitstream/handle/nsu/8917/paper_66_v2.pdf, last accessed 2019/11/05.

6. Lipuntsov, Ju.P.: Jelektronnoe gosudarstvo. Chast' 1. Modeli i arhitektura. TJeIS, Moscow (2010).

7. Nguen, M.Kh., Adzhiev, A.S. Opisanie i ispol'zovanie tezaurusov v informacionnyh sistemah, podhody i realizacija. Jelektronnye biblioteki. 7(1) (2004). https://elbib.ru/ru/article/247, last accessed 2019/11/05.

8. Neguljaev, E.A.: Universal'nye shemy metadannyh i zadacha opisanija veb- resursov. Bibliosfera. 1, pp. 43–54 (2006).

9. Polyakov, V.N.: Proekt WordNet i ego vlijanie na tehnologii komp'juternoj i kognitivnoj lingvistiki (Obzornaja stat'ja). In: Trudy Kazanskoj shkoly po komp'juternoj i kognitivnoj lingvistike TEL-2002, pp. 6–61. Kazan' (2002).

10. Fedotov, A. M., Abdeliyeva, M. N., Baidavletov, A. T., Bapanov, A. A., Sambetbayeva, M. A., Fedotova, O. A.: Conceptual model of scientific and educational information system. Vestnik NSU Series: Information Technologies 13(3), pp. 52–67 (2015).

11. Fedotov, A. M., Idrisova, I. A., Sambetbayeva, M. A., Fedotova, O. A.: Using the thesaurus in the scientific and educational information system. Vestnik NSU Series: Information Technologies 13(2), pp. 86–102 (2015).

12. 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), pp. 82–109 (2018), https://doi.org/10.25743/ICT.2018.23.5.008.

13. Heuvel, H. van den, Komen, E., Oostdijk, N.H.J.: Metadata Collection Records for Language Resources. In: Calzolari N., Choukri K., Cieri C. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC2018), May 7-12, Miyazaki, Japan, pp. 1282–1288. European Language Resources Association (ELRA), Miyazaki, Japan (2018). http://hdl.handle.net/2066/192404, last accessed 2019/11/05.

14. Kononova, O., Prokudin, D.: Synthetic Method in Interdisciplinary Terminological Landscape Research of Digital Economy. SHS Web Conf. 50, 01082 (2018), https://doi.org/10.1051/shsconf/20185001082.

15. Miller, G.A., Fellbaum, C.: WordNet then and now. Lang Resources & Evaluation. 41(2), pp. 209–214 (2007), https://doi.org/10.1007/s10579-007-9044-6.

16. 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, pp. 247–270 (2019), https://doi.org/10.1007/s10588-018-09285-y

17. Okamura, K.: Interdisciplinarity revisited: evidence for research impact and dynamism. Palgrave Communication **5,** 141 (2019), https://doi.org/10.1057/s41599-019-0352-4.
18. Raimbault, J.: Exploration of an interdisciplinary scientific landscape. Scientometrics 119(2), 617-641 (2019), https://doi.org/10.1007/s11192-019-03090-3.
19. Carusi, C., Bianchi, G.: A look at interdisciplinarity using bipartite scholar/journal networks. Scientometrics (2019), https://doi.org/10.1007/s11192-019-03309-3.
20. Giovanni, A., Ciriaco, A.D'A., Lin Zh.: A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. Journal of Informetrics. 12(4), 1182-1193 (2018), https://doi.org/10.1016/j.joi.2018.09.001.
21. Koivisto, J, Hamari, J.: The rise of motivational information systems: A review of gamification research. International Journal of Information Management. 45, 191-210 (2019), https://doi.org/10.1016/j.ijinfomgt.2018.10.013.