

International Journal
of
Advanced Statistics and IT&C
for
Economics and Life Sciences

Editor

Daniel Volovici

Lucian Blaga University of Sibiu, Romania

Managing Editor

Radu George Cretulescu

Lucian Blaga University of Sibiu, Romania

Volume 9
Number 1
June 2019

ISSN 2559 - 365X
eISSN 2067 - 354X

Editorial board

L.Rogozea, Transilvania University of Brasov, Romania

M. Breazu, Lucian Blaga University of Sibiu, Romania

D. Morariu, Lucian Blaga University of Sibiu, Romania

LIBRARY AND INFORMATION SCIENCE VIS-À-VIS WEB SCIENCE IN THE LIGHT OF THE OECD FIELDS OF SCIENCE AND TECHNOLOGY CLASSIFICATION

Marta GRABOWSKA¹,

¹ Prof. UW dr hab. , Centre for Europe, University of Warsaw, Poland

Abstract

Aims: The paper focuses on the methodological frames of Library and Information Sciences vis-à-vis Web Science in the light of the OECD Fields of Science and Technology Classification. The roots of Library Science and Information Science in Humanities and Social Sciences are described. The technological revolution which took place during and after World War II enabled the development of a new mathematics- and engineering-oriented environment for information. On this basis such new research areas like Web Science emerged. It led to a change towards an interdisciplinary character of Information Science. Method: The OECD Fields of Science and Technology Classification was analysed from the point of view of the Library and Information Science's place in this classification.

Solutions: In the OECD Fields of Science and Technology Classification Library Science has its independent place within Social Sciences while Information Science is dispersed between three main sections. It confirms the interdisciplinary character of Information Science and sets up its name as a superior covering traditional Information Science and all of new mathematics- and engineering- based research areas dealing with information. Although the name Web Science is not mentioned in this classification, we can assume that it is a sub-discipline of Information Science in the light of the OECD classification. Polish implications are mentioned.

Keywords: Library Science, Information Science, Web Science, Internet Science, the OECD Fields of Science and Technology Classification

1 Origins of Library Science, Documentatology and Information Science

If we begin our discussion on Library Science as the oldest of the three mentioned in the title – its roots will originate from Humanities and Social Sciences. This approach draws frames and indicates the main topic of its research: that is a cultural heritage of human beings as social beings registered in library documents collected, described, organized, preserved and made available for all human beings in order to disseminate its content through the communication process. Its strong and independent position has been based on its own paradigm, that is the bibliographic method understood

either traditionally like formulated by Gabriel Naudé (1600-1653) - the author of *Advice on Establishing a Library* (1627) [1] and Joachim Lelewel (1786-1861) - the author of *Two Bibliographic Books 2 vol. (1823-1826)* [2] or represented by international bibliographic standards applied today. The relationship of Library Science to other disciplines covers primarily such auxiliary fields like: history, language and literature studies, psychology or statistics. Basically, they all have its methodological roots in Humanities and Social Sciences. Thus, inductive reasoning has prevailed in the whole area of research including Library Science.[3]

The concept of culture (lat. *colere*) can be understood either narrower as a set of ideas norms and values or broadly including civilization (i.e. technological development). At present civilization is usually included into a general concept of culture. Science (knowledge) is also a part of culture defined in its broader sense but it is regarded as an autonomous part of culture of which the scope covers the sphere of application of research methods.[4]

Running a library named librarianship is a practical aspect of Library Science. Forms of documents collected in libraries have been changing and we are not going to discuss this topic here. But whatever a document has been - it has always consisted with a carrier (medium) and a human thought registered on it as an element of cultural heritage. Forms of carriers and thoughts registered this are two different things as carriers can change while thoughts can stay unchanged or *vice versa*. This approach allowed us to separate a carrier (medium) and a thought what led us subsequently from the concept of document to the concept of information (although it should be pointed out that information can't exist without a carrier). Information Science preceded by Documentation or Documentology was born on this basis especially in the sphere of the transmission of human knowledge and it is Paul Otlet (1868-1944)[5] who deserves the credit for it. Information itself is elastic, resilient, i.e. the same thought can be expressed with various codes (in different natural or artificial languages) or by applying a properly constructed metalanguage. Documentology was developed in the second part of the XIX century in relation to the fast growth of knowledge and scientific publications mainly scientific journals.[6] It turned an attention of bibliographers of that time into the mentioned autonomous part of culture i.e. science (knowledge) and a need of its fast dissemination. New forms of carriers which could fastened transmission and dissemination of human knowledge in recorded form and to process information in more flexible way became a crucial point in the development of Documentology. Still, however, the roots of Documentology and farther on of the new born Information Science originated from Humanities and Social Sciences and they applied the similar paradigm and way of thinking as Library Science. This approach to Information Science is expressed in the Tefco Saracevic's definition: "*More specifically, information science is a field of professional practice and scientific inquiry addressing the effective communication of information and information objects, particularly knowledge records, among humans in the context of social, organizational, and individual need for and use of information*". [7]

2 The technological revolution

World War II brought a deep change in technologies. The famous story on the German ciphering machine ENIGMA used by German army during the II World War [8] marked the new era in the sphere of information. Despite former methodological location of Information Science within Humanities and Social Sciences and the prevailing model of inductive reasoning being applied – a new methodological environment within mathematics, engineering and computer science enabled to perceive information as a subject of research from another angle and to build the information theory based on the deductive model of reasoning.

The circumstances of this change had no connection neither to the mentioned above methodological orientation of Library Science, Documentology and Information Science nor to researchers who had been involved in its development. Shortly before the II World War three famous Polish mathematicians and cryptologists: Henryk Zygałski (1908-1978), Marian Rejewski (1905-1980) and Jerzy Różycki (1909-1942) plus one Polish engineer Maksymilian Ciężki (1898-1951) broke the ENIGMA code. In order to break this complicated cipher mathematics-based methods were applied and not linguistic methods as it usually had been done before. This crucial change opened the door to a new approach to information. This discovery was conveyed through France to England, where British authorities created the ENIGMA deciphering center in Bletchley Park (about 60 miles in due North from London). This hidden center was active through the whole World War II never being discovered by Germans. Famous British mathematicians including Alan Turing (1912-1954) worked there during the War and it helped Aliant to win battles on Atlantic Ocean and the Battle of Britain. Specialists say that having a key to the ENIGMA code it helped to shorten the War for 2-3 years. In such circumstances, unexpectedly, information became a subject of mathematics-based research paradigm and, as it was mentioned above, it was the entirely independent process from this what was going on with Library Science, Documentology and Information Science perceived traditionally with its methodological roots in Humanities and Social Sciences.

To fasten the decipherment of messages being obtained from the secret radio watch in Bletchley Park - Alan Turing built up a machine named Colossus [9] which has evolved later on to the form which we know at present as a computer. This technology conveyed to the United States after the War led to the concept of the mathematical theory of information and the development of ICT by such scientists like Ralph Hartley (1888-1970), Claude E. Shannon (1916-2001), Warren Weaver (1894-1973), Norbert Wiener (1894-1964). This theory of information was formulated on the basis of mathematical and statistical methods including the theory of probability and stochastic (random) processes. [10]

In the *"International Encyclopedia of the Social & Behavioral Sciences we can find the following definition: "Information theory is the mathematical*

treatment of the concepts, parameters and rules governing the transmission of messages through communication systems. It was founded by Claude Shannon toward the middle of the twentieth century and had since then evolved into a vigorous branch of mathematics fostering the development of other scientific fields...[11] This approach fostered the development of such new disciplines like informatics, electronics and telecommunications and created a new environment for information whereas information itself began to operate on new types of carriers i.e. electronic carriers in the form of electronic documents. Also, new measures of information appeared like ban and decyban (Ralph Hartley and Alan Turing), entropia (C.E. Shannon) and negentropia (E.Schrödinger 1887-1961 and L. Brillouin 1889-1969) and presently used bit or shannon (from the name of C.E. Shannon who created this term) and bytes (Werner Buchholtz 1922-). The new type of communication process based on W3C standards, network and hypertext protocols led to a development of Internet - the platform on which the World Wide Web was built. All this enabled to broaden both the scope of information processed (not only knowledge but the whole universe of information) and to extend the communication process (covering the whole world). This new circumstances created so called Internet and Web Science which, however, didn't replace entirely the former Information Science in its traditional form. The question remains if we should consider Web Science and developed also Internet Science as new and independent sciences or we should include it to Information Science and to consider the last one as the interdisciplinary science.

In this point perhaps a short clarification is needed relating to Web Science *vis-à-vis* Internet Science. They both have their roots in mathematics- and engineering-based methodology although they differ by its scope. And certainly they are not the only areas which have emerged from this new environment. In fact, we are not going here to consider in depth differences between them but we can say only that Internet Science, which became a platform for the development of Web Science focuses more on technical issues like hardware design, access technologies and content delivery while Web Science emphasizes more on its semantic, information retrieval and social networking issues. More details on this topic can be found in relation to the European Union's ENIS [12] project, from the paper *A Disciplinary Analysis of Internet Science*" by Clare J. Hooper, Bruna Neves and Georgeta Bordea [13].

If we consider Information Science as an interdisciplinary science undoubtedly Harold Borko's definition would be the most accurate. He formulates it as follows: *"Information Science is that discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information. This includes the investigation of information representations in both natural and artificial systems, the use of codes for efficient message transmission, and the study of information*

processing devices and techniques such as computers and their programming systems. It is an interdisciplinary science derived from and related to such fields as mathematics, logic, linguistics, psychology, computer technology, operations research, the graphic arts, communications, management, and other similar fields. It has both a pure science component, which inquiries into the subject without regard to its application, and an applied science component, which develops services and products." [14] Such approach to Information Science as Harold Borko defined we can find also in the *OECD Fields of Science and Technology Classification*.

3 Library and Information Science in the light of the OECD Fields of Science and Technology Classification

In relation to the reform of the higher education sector carried out during last two years in Poland, the new bill was laid down by the Polish Parliament in 2018 (*Higher Education and Science Act*) [15]. Within this reform the *OECD* (The Organization for Economic Co-operation and Development) *Fields of Science and Technology Classification* is recommended as the key scheme for structural changes in universities and other higher research and education bodies in order to strengthen mutual relationships between research, education and economy. This change has an impact on a place of Library and Information Science in higher education units in the country.

The *OECD Fields of Science and Technology Classification* [16] consists with five basic sections:

1. Natural Sciences
2. Engineering and Technology
3. Medical and Health Sciences
4. Agricultural Sciences
- 5. Social Sciences**
6. Humanities

Within the section nr 5 Social Sciences the following disciplines are enumerated:

- 5.1. Psychology
- 5.2. Economics and business
- 5.3. Educational sciences
- 5.4. Sociology
- 5.5. Law

5.6.Political Science

5.7.Social and economic geography

5.8.Media and Communications

5.9.Other social sciences

Within the section **5.8 Media and Communication** there are:

5.8.a Journalism

5.8.b Information Science (social aspects)

5.8.c Library Science

5.8.d Media and Communications

Both Library Science and Information Science are included in this section, however, Library Science is perceived here as one of the Social Sciences (not Humanities any more) and is enumerated separately and perceived as an independent from Information Science certainly considering its strong methodological basis (bibliography, bibliographic standards) and well defined area of practice (library). What relates Information Science (perceived here as one of the Social Sciences, too) only its social aspects are mentioned here suggesting another location for its mathematics- and engineering- based part. This other part of Information Science we can find in the general section nr 1. Natural Sciences.

1. Natural Sciences

1.1.Mathematics

1.2.Computer science and informatics (information science)

1.2.a Computer science (algorithms), informatics (information science) and bioinformatics (computer equipment belongs to 2.2. while social aspects belong to 5.8)

1.3.Physical sciences

1.4.Chemical sciences

1.5.Earth and related environmental sciences

1.6.Biological Sciences

1.7.Other natural sciences

The location of the other part of Information Science in sections 1.2.and 1.2.a shows clearly its roots in mathematics. Additional reference mentioned in point nr 2.1.a to section 2.2. of the *OECD Classification* let us farther on through the general section nr **2. Engineering and Technology** which consists with the following areas:

2. 1.Civil engineering

2. 2.Electrical engineering, electronics, computer engineering

2. 3.Mechanical engineering

- 2. 4. Chemical engineering
- 2. 5. Materials engineering
- 2. 6. Medical engineering
- 2. 7. Environmental engineering
- 2. 8. Environmental biotechnology
- 2. 9. Industrial biotechnology
- 2.10. Nanotechnology

to the section nr **2.2. The Electrical engineering, electronics, computer engineering**, which covers six areas each being a subject of interest of this part of Information Science:

- 2.2.a Electrical engineering and electronics
- 2.2.b Robotics and automation
- 2.2.c Automation control systems
- 2.2.d Engineering and communications systems
- 2.2.e Telecommunications
- 2.2.f Computer equipment and computer architecture

From this what was mentioned above we can assume that according to the *OECD Fields of Sciences and Technology Classification*, Library Science and social aspects of Information Science belong to Social Sciences. Information Science is perceived here as an interdisciplinary science covering under this name also its mathematics-based and engineering-oriented part. Web Science (and Internet Science) emerged from mathematical theory of information but it is being strongly pointed out by such famous researches as Tim Berners-Lee or Ben Schneiderman that Web Science (and Internet Science) are not able to exist and function properly as an independent sciences without of its Social Sciences support.

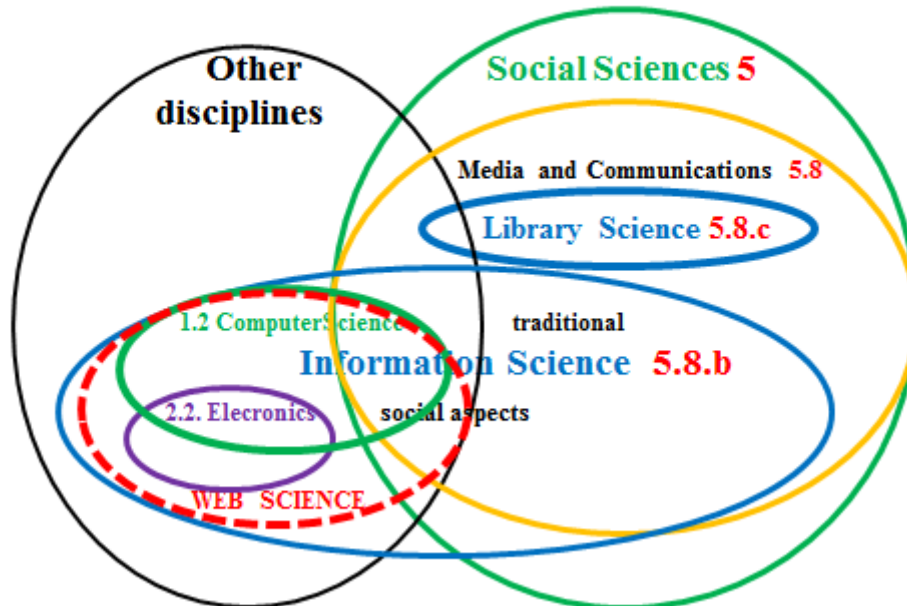
Tim Berners-Lee says: "*Web Science must coordinate engineering with a social agenda, policy with technical constraints and possibilities, analysis with synthesis – it is inherently*

interdisciplinary.... ...engineering needs to go hand in hand with a social process of negotiation" [17]. Ben Shneiderman says: "*Web Science is processing the information available on the Web in similar terms to those applied to natural environment...*" [18]

Thus, in case of Web Science (and Internet Science) the story begins from another angle, i.e. from information theory but requires also the methodological support of Social Sciences without which it can't function properly.

Below the project of the scheme of Library Science, Information Science and Web Science according to the *OECD Fields of Science and Technology Classification* is presented.

**Library Science and Information Science vis-à-vis Web Science
in the light of the *OECD Fields of Science and Technology Classification***



Source: The scheme created by the Author on the basis of the *OECD Fields of Science and Technology Classification*

Conclusions:

According to the *OECD Fields of Science and Technology Classification* there are two separate sciences: Library Science and Information Science. Library Science relies on its own paradigm as one of the Social Sciences (and certainly Information Science is one of its auxiliary sciences). Information Science is dispersed between Social Sciences, Natural Sciences (mathematics) and Engineering and Technology. The first part, earliest of the three, operates on the basis of the methodology of Social Sciences while the last two parts emerged later on being built on the basis of the mathematical theory of information. While the first part can still function independently ("traditional" Information Science) being also a support for the other two parts, the last two parts still can't exist without a support of the first one. The question remains, of course, if and when the last two parts of Information Science would begin to operate without a support of the first one? It seems, that the autonomous robots are already the outpost of this phenomenon...

Meanwhile, the *OECD Classification* confirms the interdisciplinary character of Information Science and its superior name for the whole research area of information having regard presumably on its historical priority. Although,

names Web Science and Internet Science are not mentioned in the *OECD Classification* but *per se* we can consider them as its sub-disciplines.

References:

- [1] Ajdukiewicz K., *Zarys logiki*. Warszawa: Państwowe Zakłady Wydawnictw Szkolnych, 1960
- [2] Barker Ch., *Cultural Studies. Theory and Practice*. London: Sage Publications, 2003
- [3] Barnes-Lee T., Hall W., Hendler J.A., *A Framework for Web Science*. Boston Delft: Now Publishers Inc. 2006 s. 3-4 [online] [access: 20 June 2019] https://books.google.pl/books?id=VIC4pjOu2vwC&printsec=frontcover&hl=pl&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
- [4] Borko H., *Information Science. What is it?* "American Documentation" 19 (1), 3-5 1968 p. 3 [online] [access: 20 June 2019] <https://pl.scribd.com/document/81459529/Borko-H-Information-Science-What-is-It>
- [5] Chirikjian G. S., *Stochastic models, Information Theory, and Lie Groups*. Vol. 1. Boston, Basel, Berlin: Birkhäuser, 2009
- [6] Copeland E.J., *Colossus: the secrets of Bletchley Park's codebreaking computers*. Oxford: Oxford University Press cop., 2010
- [7] Dembowska M., *Nauka o informacji naukowej (Informatologia). Organizacja i problematyka badań w Polsce*. Warszawa: Inst. Inf. Nauk. Tech. i Ekonom. 1991 (Seria: *Informacja Naukowa*)
- [8] *Fields of Science and Technology* [online][access: 20 June 2019] https://en.wikipedia.org/wiki/Fields_of_Science_and_Technology
- [9] Hooper, C.J., Neves B., Bordea G., *A Disciplinary Analysis of Internet Science*. [In:] Tiropanis T., Vakali A., Sartori L., Burnap P. (Eds.). *Internet Science. Second International Conference INSCI 2015. Brussels, Belgium, May 27-29, 2015. Proceedings*. Springer International Publishing AG, Switzerland (Springer-Science-Business Media), 2015 s.63-77 (*Lecture Notes in Computer Science* 9089)
- [10] Inglis F., *Culture*. Cambridge: Polity Press, 2004
- [11] Lelewel J., Bandtkie J. S., *Bibliograficznych ksiąg dwoje*. Warszawa: Wydawnictwa Artystyczne i Filmowe, 1980
- [12] Martignon L., *Information Theory*. [In:] *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2001 p. 7476-7480 [online] [access: 20 June 2019] <https://www.sciencedirect.com/science/article/pii/B0080430767006082>
- [13] Narodowe Centrum Badań i Rozwoju, *Wykaz dziedzin nauki i techniki według OECD* [online] [access: 20.06.2019]

https://www.ncbr.gov.pl/fileadmin/user_upload/import/tt_content/files/2_wykaz_dziedzin_nauki_i_tech_nik_wedlug_klasyfikacji_oecd.pdf

- [14] Naudé G., Taylor A., *Advice on establishing a library*. Berkeley: University of California Press, 1978
- [15] *Network of Excellence in Internet Science*. [In:] European Commission. Cordis. EU research results [online] [access: 20 June 2019] <https://cordis.europa.eu/project/rcn/101725/factsheet/en>
- [16] Otlet P., *Traité de documentation: livre sur le livre: théorie et pratique*. Liege: Centre de lecture publique de la Communauté française de Belgique, 1989
- [17] Rakus-Andersson, E., *The Polish Brains Behind the Breaking of the Enigma Code Before and During the Second World War*. Berlin-Heidelberg-New York: Springer-verlag, 2003
- [18] Saracevic T., *Information Science*. [In:] Bates m.J., Maack M.N. (Eds.) *Encyclopedia of Library and Information Science*. New York: Taylor & Francis. pp. 2570-258M. [online] [access 20 June 2019] <https://tefkos.comminfo.rutgers.edu/SaracevicInformationScienceELIS2009.pdf>
- [19] Schneiderman B., *Web Science*. „Communications of the ACM” 2007 50 (6) s.25
- [20] *Ustawa z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce*. Dz. U. 2018 poz. 1668

STUDY ON THE MAPPING OF RESEARCH DATA IN THE REPUBLIC OF MOLDOVA IN THE CONTEXT OF OPEN SCIENCE

Nelly Turcan¹, Andrei Rusu², Rodica Cujba³,

¹ Univ. Prof., Information Society Development Institute, Moldovan State University; Moldova

² Assoc. Prof., Institute of Mathematics and Computer Science „Vladimir Andrunachievici”, Moldova

³ Sc. Res., Information Society Development Institute, Technical University of Moldova, Moldova

Abstract

The Open Science concept represents a new approach to the way in which scientific research based on cooperation and new ways of knowledge dissemination is carried out and organized, using new digital technologies, new tools for collaboration, and R&D infrastructure to ensure open access to research data.

This study uses data collected in May - July 2018 within a survey that aimed at investigating the scientific data ecosystem in the Republic of Moldova. Findings show that, although there are some concerns about the loss of property rights and copyright infringement in case of sharing and open access to research data, Moldovan academia is ready to provide access to research data. The research has highlighted that a new challenge is needed to solve scientific data issues by creating new type of infrastructure to ensure data retention, broad access to research results for the purpose of their dissemination and use, and creating new research opportunities based on research data.

Keywords: open science, open research data, e-infrastructure, Republic of Moldova

1 Introduction

The speed of progress in science has always been dependent on how effectively scientists can communicate their results to colleagues but also to those who want to implement these results in new technologies and practices.

Currently, we are witnessing an important change in what we call science in terms of organizing, conducting, evaluating, using and disseminating research results. This change that reflects the term “Open Science” is determined by the development of new technologies, increasing social role of scientific

research, the current political and institutional context (Cuciureanu, 2018, p. 15).

According to European Commission Open Science is about the way research is carried out, disseminated, deployed and transformed by digital tools, networks and media (European Commission, 2018). The Organization for Economic Co-operation and Development mentions that Open science commonly "refers to efforts by researchers, governments, research funding agencies or the scientific community itself to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction as a means for accelerating research; these efforts are in the interest of enhancing transparency and collaboration, and fostering innovation" (OECD, 2015, p. 7).

Open Science brings social, economic, cultural, political and technological change, based on openness and connectivity, on how research is designed, performed, used, assessed, and preserved. Open access platforms, open infrastructures, open data tools, open educational resources, open evaluation methods, open collaboration, or open citizen science activities are irreversible trends that are impacting all scientific actors and have the potential to accelerate the research cycle (Vicente-Saez and Martinez-Fuentes, 2018). By increasing access to publications and data, universities, research institutions, firms and individuals may use and re-use scientific outputs to produce new products and services.

One of the key elements of open science is open access to research data. Open research data is research data that combines the characteristics of open data and the types of research data (Onyancha, 2016).

As World Bank mentioned in his report (World Bank, 2018, ch.1) data is considered the new gold, or the new oil, and like oil, unprocessed data has relatively little value and needs to be mined, refined, stored, and sold on to create value.

In a research context, there is a growing opinion that most research data should be open, particularly data from publicly funded projects. This point of view is driven especially by research funder requirements for sharing and re-use data, upon principles regarding research data such as to be findable, accessible, interoperable and re-usable (FAIR principles). The research data are made open for two purposes: to provide evidence that the research was conducted properly and to provide data for reuse and the generation of further findings and outputs (Childs et al., 2014).

The updated Directive on Open Data and Public Sector Information obliges Member States to "support the availability of research data by adopting national policies and relevant actions aiming at making publicly funded research data openly available ('open access policies') following the principle of open by default and compatible with FAIR principles" (European Parliament, 2019).

Scientists are particularly interested in data collection, and the success of each experiment is determined by the new data generated, which can contribute to

advancing scientific knowledge. Any scientific research involves performing an observation, generating a hypothesis, running an experiment, and collecting data. Traditionally, for any research, the amount of data collected by scientists was not very extensive, and its analysis did not require the use of technology. Previously, scientists used technology in a very limited way, and data evaluation was not done using algorithms or software. However, significant changes have taken place over the past two decades, and changes in software and tools have made data acquisition and analysis a very important part of research.

2 Opening science in the Republic of Moldova

At present, in many countries, there are various actions to promote the transparency of governance; promoting dialogue between governance and citizens; the use of new technologies in governance and in dialogue with citizens. Opening and sharing public data and information is an essential policy in building a more open, responsible and efficient government. Open, democratic and transparent science could be a major factor for any government in the progress of a country.

Opening science requires a new systemic approach, especially in nationally and internationally agreed strategies and policies.

Special strategies and policies dedicated to Open Science are not yet developed in the Republic of Moldova. According to the national legislation and the number of open access policies approved at national and institutional level, the Republic of Moldova cannot be called a politically open territory. However, there are some encouraging examples, as well as a positive forecast for the future to support open access and open science in the Republic of Moldova.

Promoting and implementing of Open Science policies at national level is necessary for at least three reasons (Gh. Cuciureanu et al., 2018, p. 16):

- a) the transition to Open Science is an official policy of the European Union and the Republic of Moldova declared its intention to integrate into the European Research Area;
- b) the Open Science concept radically changes the way to do research, and its lack of implementation in the Republic of Moldova will make non-competitive native science;
- c) the legislation of the Republic of Moldova provides for certain elements of Open Science (even if not consolidated) that have to be implemented.

In the Republic of Moldova, the state policy in the field of science and innovation is carried out under the Code on Science and Innovation of the Republic of Moldova (Parliament of the Republic of Moldova, 2004). Several articles of the Code assure guaranteed access to scientific information. In accordance with the current legislation, the State guarantees: support through access to information, through its dissemination; information assurance of the

topics from the scientific and innovation field; free and non-discriminatory access to scientific-technological information resources. However, at state level open access to scientific research is not confirmed in the Republic of Moldova, primarily that financed from the public money, and besides this, the mechanisms for ensuring free and open access to the scientific and technological information resources and open access to research data are not specified.

In order to support open access and open research data there have been launched several projects in the Republic of Moldova. This project focuses on improving the quality of academic studies which also focus on open and free access to scientific information and data.

3 Mapping of research data in the Republic of Moldova

In order to map the situation regarding generation, gathering, use, sharing and preservation of research data obtained within research projects carried out in the Republic of Moldova, Information Society Development Institute conducted a survey in May-July 2018.

The survey on the mapping of the research data ecosystem in the Republic of Moldova was carried out within the framework of the project *Development of the conceptual and methodological framework for data e-Infrastructure in the field of research, development and innovation in the Republic of Moldova* (e-IDSM) <https://idsi.md/en/e-idsm>). Unlike the previous survey this one was focused exclusively on research data. The main goal of this survey was to identify the needs of the RDI community in the Republic of Moldova on the management of scientific data over their lifecycle (creation, processing / analysis, storage / preservation, sharing / access and use). The specific objectives of the survey were:

- to identify the types / formats and sources of research data;
- to find out the modes of storing and preservation of research data;
- to discover the ways research data are processed and analyzed;
- to learn the procedures of research data management;
- to determine methods of sharing, access and use of research data.

3.1 Methodology

The survey covered questions regarding the entire lifecycle of research data. It included five sections:

- Creation and storage of research data;
- Storage and preservation of research data;

- Data processing and analysis;
- Data sharing;
- Research data management.

Responses were collected from 48 RDI institutions (92% success rate), including 13 higher education institutions. Respondents with various positions within these institutions participated in the survey, including: 25 heads of RDI institutions (12.3%); 42 project managers (20.7%), 65 laboratory / research group managers (32%), 34 scientists (16.7%), 23 university teachers (11.3%), 4 PhD students (2%), other positions (5%).

The authors were members of the research projects' teams and have participated in the design of the surveys, collection and aggregation of the results.

3.2 Results and Discussions

The data obtained in this survey allows us to find out that Moldovan academia produces and generates different kinds of research data. Most respondents of the survey mentioned that they produce or generate the following types of data for research: text (86.2%), images (66%), numerical data (62.6%), tabular data (55.7%) (Figure 1). Only 13.8% of respondents do not produce or generate any type of research data. Other types of data were also mentioned, including: DICOM images, archive quotes, technical drawings, protocols, algorithms, programs, maps, etc.

Concerning the format of the data generated or collected, the survey participants indicated that they use different formats for scientific data (Figure 2). Respondents have specified that they use all categories of formats specified in the questionnaire. However, research data is mainly generated and collected in the following formats: text (93.1%), presentations (83.2%), graphics (67%), calculation sheets (46.8%), data bases (43.3%) and statistical software files (25.6%).

Taking into account data formats used by researchers, as well as the needs described by them in the survey, Moldovan researchers can be divided into two categories:

- Researchers using relatively widespread software tools in the academic and research environment, such as: Microsoft Office, SPSS, Adobe FineReader, and others.
- Researchers using research-specific software solutions such as: ArGIS, Geoportal, Mathematica, FoxPro, Endnote, 1C, EViews, GAMESS, Gaussian09 and others.

Scientific data can be obtained or generated as a result of research activities as well as from various sources. Survey participants noted in particular the following ways of production and generation research data:

- results of the experiments (69.3%),
- results of the observations (72.3%),
- scientific publications (72.8%), statistical sources (52%),
- survey results (33.7%).

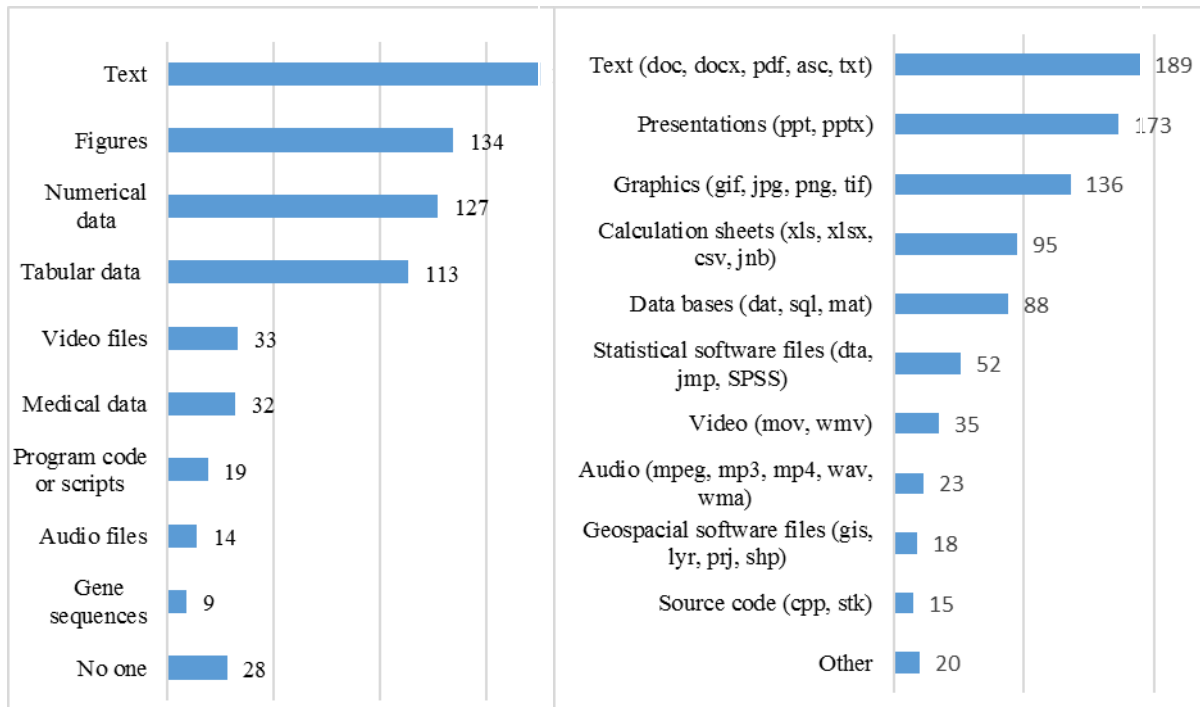


Figure 1. Types of digital data produced / generated for research

Figure 2. Types of generated / collected data formats

For some areas of research, data can be obtained from social media, sensor data, archive files, demographic forecasts, bilateral and international projects, media, weather station data, medical images, etc. (Figure 3).

Sources of input data used in the research process largely coincide with the mode in which scientific data are produced and generated (Figure 4). Thus, experiment data (74.3%) is the main source for the research data collection. Also, important sources of extraction, collection and input of research data are data from studies and surveys (66.3%), official national statistics (48.5%), international statistics (45%), public data sets (35.1%). At the same time, respondents noted analytical / theoretical results, archive data, social media, mass media, clinical data as inputs of data used in the research process etc.

Procedures for storing and archiving research data are very important not only for long-term preservation of research results but also for ensuring the integrity of these data. Thus, 102 (50.2%) of the respondents mentioned that they take steps to preserve research data, 61 (30%) of respondents said they did not take measures to preserve the data, and 40 (19.7%) of respondents

said that they do not know if data preservation measures are under way in the respective research projects. Among the specified methods of storing research data, the researchers indicated two large categories (Figure 5):

- locally, on the personal computer (96.6%) and / or on physical support, i.e. on paper (72.9%), CD, USB, or external HDD (63.5%) etc.;
- online in databases, specialized institutional repositories or research laboratories (24.6%), the institutional computer network (32%) and / or internationally, Dropbox, Google Drive (30.5%), etc.

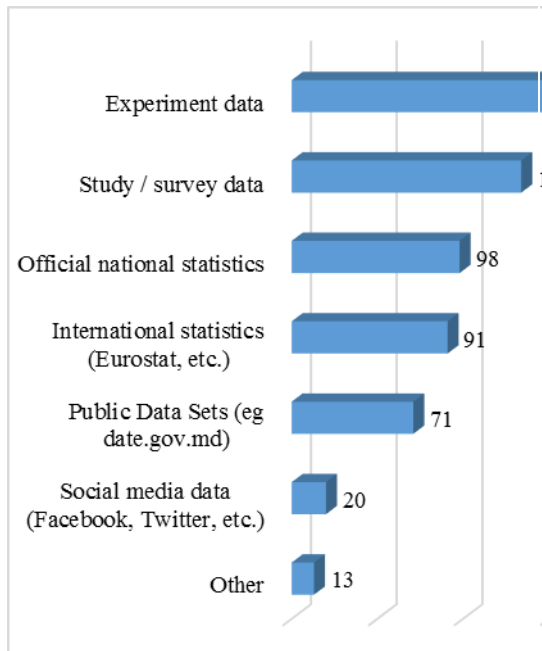


Figure 3. Sources of input data used in the research process

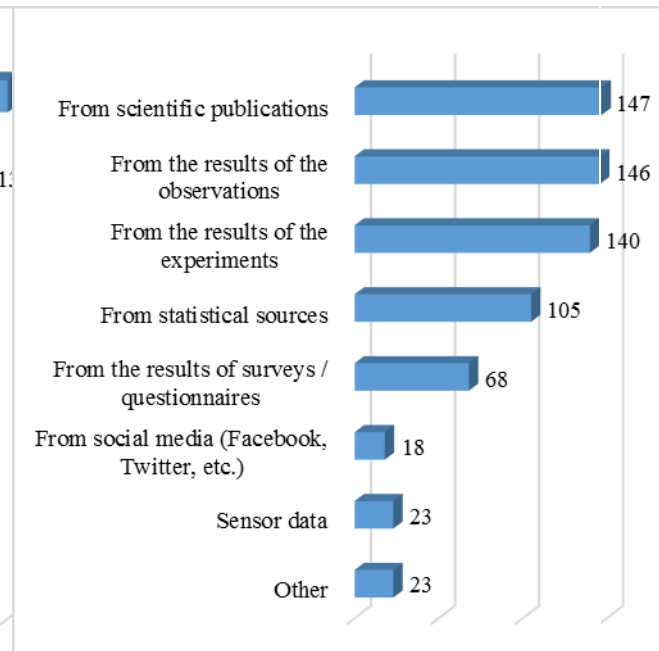


Figure 4. The ways the scientific data is obtained (produced / generated)

It should be noted that many respondents have specified several ways of preserving research data, which contributes to the safety of data retention. However, 30% of respondents do not take measures to preserve research data, and 19.7% do not know if the institution or laboratory is taking such measures.

Respondents noted that they protect research data by limited access to research data (66%), password (63.1%), backups (34%), data anonymization (22.2%), data coding (11.8%), etc. More than 10% of respondents do not protect research data, and 3% of them destroy data after use.

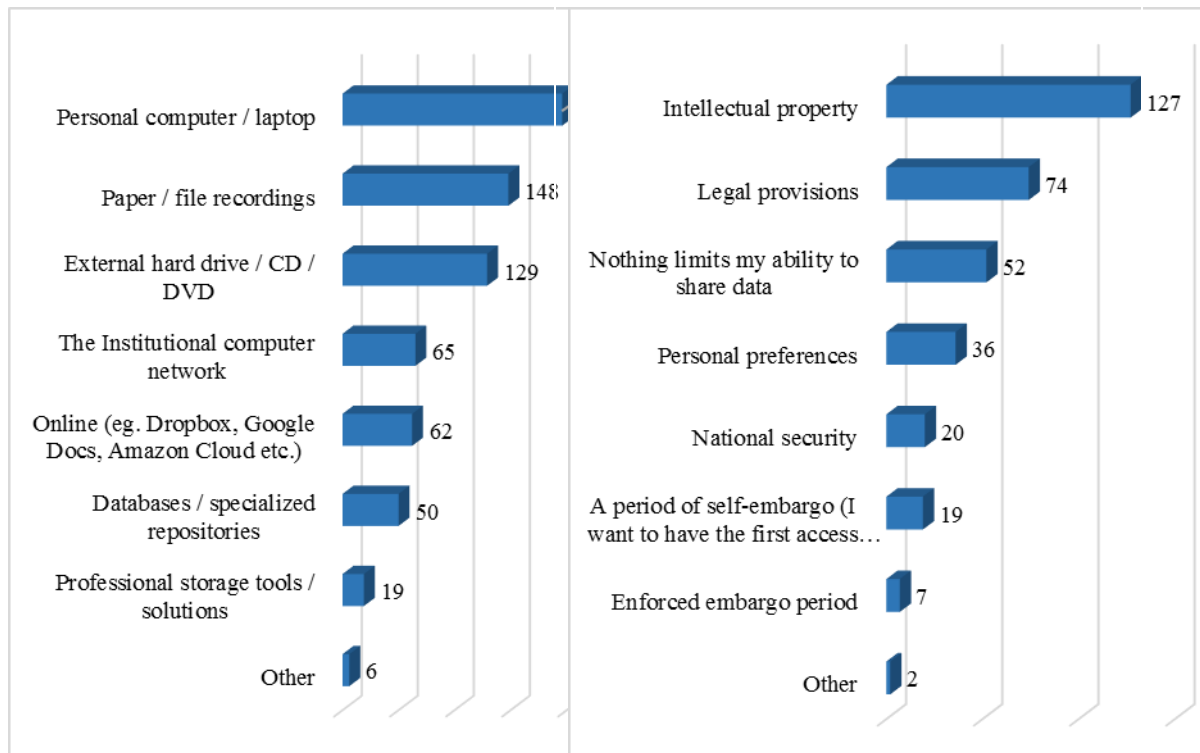


Figure 5. Methods of storing research data

Figure 6. Restrictions that limit the ability to share data

Currently, when scientific research is highly interdisciplinary and collaborative, it is necessary to share and use data interoperability procedures. Thus, 61 (31%) of the respondents replied that they share research data, while 122 (60.1%) share them according to the research project policy. Researchers also noted that there are some restrictions limiting the sharing of research data (Figure 6), such as intellectual property (62.6%), legal provisions (36.5%), national security (9.9%), embargo or self-embargo (12.8%). About one-fourth of respondents consider that nothing limits data sharing (25.6%), and 17.7% of those surveyed noted that data sharing limits depend on personal preferences.

Research data is usually documented using metadata. 76 respondents (37.4%) noted that they document or record certain metadata for scientific data or data sets, while the majority of respondents (62.6%) do not. Only 35 respondents (21.1%) use metadata standards when recording or documenting research data. Among the metadata standards used were mentioned CERIF, Diagnostic Classifier CIM10/HL7, Archival Standards, BibTeX, standards of National Bureau of Statistics, ISO 28258, ISO 11074 and ISO 15903, GenBank, CNAS, EUROSTAT, EIOPA, DICOM, etc.

It is important to have Data Management Plans for institutional data management policies or procedures as well as for research projects. The results of the survey revealed that 107 respondents, which make up more than 50% of the survey participants, do not know or believe that there are no institutional policies and procedures regarding the management of research

data (Figure 7). With regard to the development of the data management plan for research projects, only 21 respondents (10.3%) stated that funding agencies had requested such a plan (Figure 8).

Only 96 respondents (47.3%) know about the existence of institutional data management policies and procedures. They noted that there are various policies and procedures in place to protect, store, archive, share research data, among which: privacy policy, data storage policy, institutional policy on open access, institutional policy on intellectual property and technology transfer, primary data verification policy, old data removal policy, strict journaling of records, experiments and tests registries, contracts with organizations, non-disclosure agreements, internal networks specifying data access rights, etc.

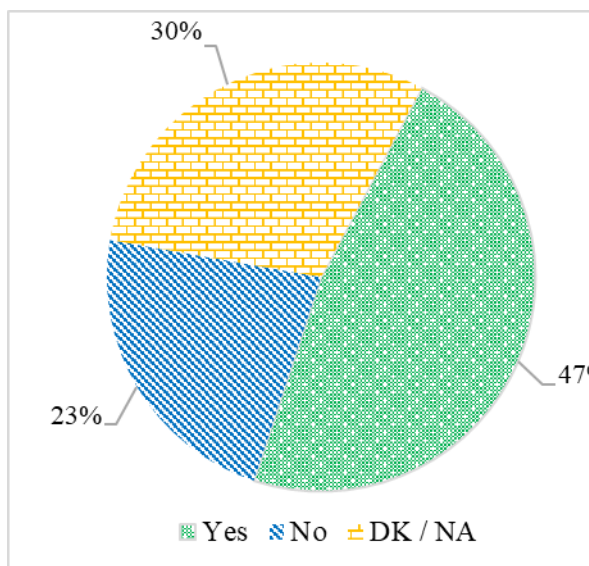


Figure 7. Existence of institutional management data policies and procedures

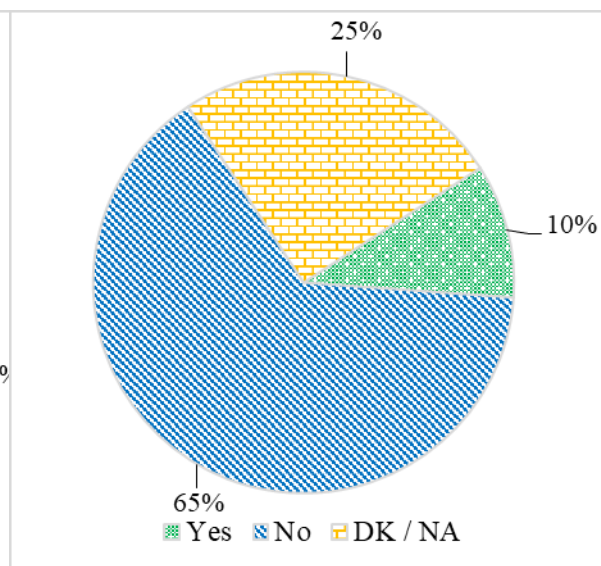


Figure 8. Request from financial agencies of Data Management Plan

Respondents noted that there are a number of issues they face in managing research data. The most common problems are:

- non-standard file formats with processing difficulties (34,5%),
- no financial resources are available to ensure data management (33%),
- difficulties to find files that have been developed themselves or by their colleagues, i.e. file versions, map structure, etc. (28,6%),
- searching for or accessing the scientific data of former colleagues, i.e. former PhD student, ex-employee (26,6%),
- identifying the storage location of data files, i.e. external hard drive, USB, DVD / CD, network (19,7%),

- there is no institution-based data management tools or solutions (19,2%), security and data file protection (15,8%),
- problems with setting up the data owner (11,3%),
- legal issues as a result of international data transfer (7,4%) etc.

However, it should be noted that the majority of survey participants (170 respondents – 83.7%) believes that training on research data management is needed (Figure 9). Respondents emphasized the necessity of training researchers from different fields on research data management technologies.

The survey highlighted the need to improve the circulation of knowledge and access to research data. Thus, 31% of survey participants noted that they unconditionally share research data, 60.1% said they share the research data according to the conditions specified in the research project, only 14.3% do not share their research data.

Survey participants were asked what they would choose in case they would share or plan to share research data. Most respondents noted that they will present data to journal as a support for the publication (64%). Also, other data sharing options have been identified, such as sharing the research data to colleagues on request or informally (39,9%), the data will be available online on the project or institution site (34,5%), storing data in a specialized database or repository (33,5%) etc.

This survey provided questions on open access to research data resulting from public funding (Figure 10). The majority of researchers (177 respondents) believe it is necessary to open the data resulting from public money-funded research. 57 respondents (28.2%) opted for unconditional open access, 29 respondents (14.4) were granted open access after an embargo, and access under certain conditions was supported by 91 participants' survey (45%). Only 25 researchers (12.4%) do not support the opening of data from state funded research.

At the same time, the survey participants have specified that research data must be accessible to colleagues, scientific community, PhD students, decision-makers, educational institutions and other users, and one of the primary conditions for using research data is to cite the source. Also, it was mentioned that there is no mechanism for managing and coordinating international projects in the Republic of Moldova, some data banks are not accessible to the public, and researchers do not have sufficient skills in managing research data.

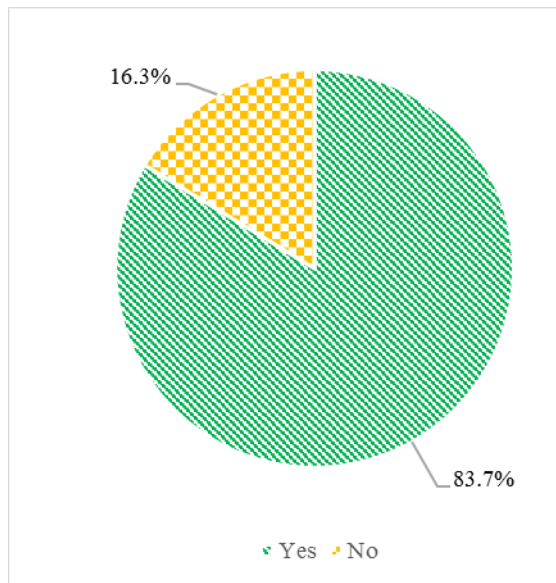


Figure 9. Moldovan academia's opinion on the necessity of the training on research data management

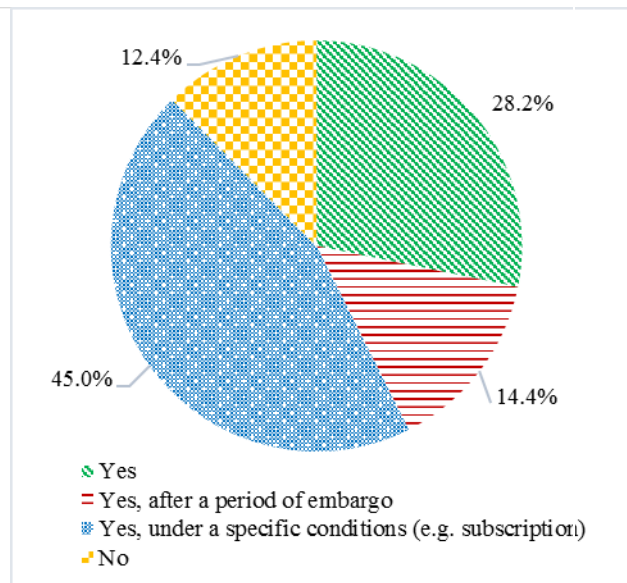


Figure 10. Open Access to research data resulting from public funding

4 Conclusions and Recommendations

In conclusion, we can mention that Moldovan academia is ready to provide access to research data. Most of researchers use digital media to access research data, but prefer to use traditional data storage formats (Word, Excel, PDF, etc.). Only some researchers use modern and innovative tools to process, access, store and archive research data. One of the main issues that discourages research data sharing is the issue of copyright protection. There are some concerns about the loss of property rights and copyright infringement in case of sharing and open access to research data. At the same time, in RDI institutions of the Republic of Moldova, the management of research data management is not implemented. There are problems related to long-term preservation, storage, sharing and access to research data.

Concluding results of these studies, the following recommendations can be made:

- Continuous analysis of the tendencies in research data management.
- Analysis of the international experience in the field of research data management.
- Establishment and approval of policies on research data management in research projects and / or research institutions.
- Training and familiarization of Moldovan academia in different fields of research data management.

- Training researchers in the Republic of Moldova on open source solutions that could be used in the research process as well as in the special case of research data management.
- Establishing rules / procedures / customs for research data management to be known to researchers, to be adopted by researchers and addressing all stages of research data management.
- Organization of round tables or other activities for presentation of solutions used in the field, as well as for exchange of views and experience in the field of research data management.

Acknowledgments

The study was performed due to the support of the research project 18.50.07.10A/PS "Elaboration of conceptual and methodological framework for e-Infrastructure of data in the RDI field of the Republic of Moldova (E-IDSM)" (2018-2019) funded by the NARDRM.

References

- [1] European Commission. *Open Science*, Last update: 24 April 2018. Available at: <https://ec.europa.eu/digital-single-market/en/open-science>
- [2] European Parliament. *European Parliament legislative resolution of 4 April 2019 on the proposal for a directive of the European Parliament and of the Council on the re-use of public sector information*, Brussels, 4 April, 2019. Available at: www.europarl.europa.eu/doceo/document/TA-8-2019-0352_EN.pdf
- [3] Gh. Cuciureanu et al. *Open Science in the Republic of Moldova: Study [Știința Deschisă în Republica Moldova: Studiu] (in Romanian)*. Institutul de Dezvoltarea Societății Informaționale, Chișinău, 2018.
- [4] O.B. Onyancha. "Open Research Data in Sub-Saharan Africa: A Bibliometric Study Using the Data Citation Index". *Publishing Research Quarterly*, 32(3): 227-246, 2016. Available at: <https://doi.org/10.1007/s12109-016-9463-6>.
- [5] OECD. "Making Open Science a Reality. OECD Science". *Technology and Industry Policy Papers*, 25. Publishing, Paris, 2015. Available at: <http://dx.doi.org/10.1787/23074957>.
- [6] Parliament of the Republic of Moldova. *The Code on Science and Innovation of the Republic of Moldova [Codul cu privire la știință și inovare al Republicii Moldova] (in Romanian)*. Nr. 259, 15.07.2004. Available at: <http://lex.justice.md/md/286236/>.
- [7] R. Vicente-Saez and C. Martinez-Fuentes. "Open Science now: A systematic literature review for an integrated definition". *Journal of Business Research*, 88: 428-436, 2018. Available at: <https://doi.org/10.1016/j.jbusres.2017.12.043>.
- [8] S. Childs, J. McLeod CHILDS, E. Lomas, G. Cook. "Opening research data: issues and opportunities". *Records Management Journal*, 24(2): 142-162, 2014. Available at: <https://doi.org/10.1108/RMJ-01-2014-0005>.
- [9] World Bank. *Information and Communications for Development 2018: Data-Driven Development*, 2018. Available at: <https://doi.org/10.1596/978-1-4648-1325-2>.

DIGITAL RISKS. CASE STUDY ON DIGITIZATION PROJECTS OF THE LBUS LIBRARY

Rodica-Maria Volovici¹, Elena Mărginean², Ioan-Irinel Vișa³,

¹ Lect. univ. dr., Library of the "Lucian Blaga" University of Sibiu, Romania

² Librarian, Library of the "Lucian Blaga" University of Sibiu, Romania

³ Library IT coordinator, Library of the "Lucian Blaga" University of Sibiu, Romania

Abstract

Digital technologies changed all the social life domains in society, as we are now living in the digital era, in the information society or in an interconnected world. Besides the improvements in every day life, digital changes also brought inherent risks, from cyber-security, hacking, cyber-bullying, to the vulnerability of personal data, or the mental health consequences of information explosion.

This article reviews the risks of the changes brought by the digital transformation on libraries in general, with examples of the LBUS Library, drawing from the last 10 years of development experience of informatic book management systems (electronic catalogue) and of the digital library system. A group of library experts took part in local cultural projects promoting the city of Sibiu, as well as in four major European projects focusing on "Europeana", and developing highly valuable cultural, historical, and scientific digital collections. As relatively new and highly complex technical activities comprising a high volume of new information, the management of these projects also posed risks related to decision-making and to choosing the best solutions to ensure their success. We have identified and highlighted the major risks.

Keywords: digital society, digitization risks, library management, digital library

1 Premises: Risks in traditional libraries versus risks in modern, digitized libraries

The risk concept has numerous meanings; after a detailed analysis, we ascertain that risk is mostly associated to negative factors: uncertainties, insecurities, probabilities, loss, damages that lead to the production of undesirable consequences.

Risks must be identified and evaluated according to the probability of potential events and the possible impact of the occurrence of that possibility[1].

When we look at things from a different perspective, risk appears when the organization is unable to meet its predetermined objectives [2]. In this case, risks represent a threat that can only be identified and defined in relation to the objectives and in direct correlation with the mission of the organization.

In project management, the risk is used to measure the probability and the effect of an event on the objectives of a project [3], while digital risk is defined as exposure to the loss or the partial/total destruction of the digital data needed to fulfil the organization's objectives [4].

When performing a "digital risk" evaluation at the level of a digitized organization, one must consider and analyze multiple aspects related to atomized processes, the architecture and the complexity of IT systems, the quantity and the quality of data/metadata, as well as the level of organization and the institutional management, and cultural aspects [5].

Digital transformations have had a major impact on libraries as well; their high level of automation and digitization came with IT processes through which the information and the data on physical formats was transferred to an online digital environment, so that they can be accessed by more users anywhere and at any time.

Therefore, digitization can be seen as a great opportunity: library services have improved considerably, as users (pupils, students, professors, researchers) can access information much easier and much quicker. Changes have also been implemented at the level of the library's management and organization, through the automation brought about by IT systems.

However, the same digitization, through its key elements (computer network, the volume of data/metadata, IT systems), has brought specific risks[6], besides the traditional ones.

A February 2019 OECD study entitled "How's Life in the Digital Age? *Opportunities and Risks of the Digital Transformation for People's Well-being*" [7] lists the major risks of the digital society:

The **digital divide** refers to the differences regarding the access to and the use of the internet and the digital competences (Dictionary: "*there is a 'digital divide' between rich and poor in terms of computer use*");

The lack of "digital literacy" equals the lack of knowledge and of emotional skills to sort through/choose quality information, to self-control in digital involvement and to avoiding mental health issues;

Digital insecurity. personal/private data, cyber-hacking, cyber-bullying.

The difference in technological development, and implicitly the differences in access to the internet have led to an even higher discrepancy in the world, as this gap between access and lack of access to information represents a major social risk for the contemporary world.

This is connected to the lack of digital literacy, of training regarding the correct use of information means: computers, tablets, software, the internet. This results in the need to integrate "digital training" in education, in organizations, in society, which is so necessary to prepare users who are used to credit cards, e-mail and online accounts to deal with cyber-attacks, data theft, fraudulent use of personal data.

Moreover, if we talk about the information explosion brought about by the internet, with possibilities for free speech and to report events from various standpoints, which also leads to misrepresentation, the so-called fake news that has become omnipresent, the library is a source of verified information, organized knowledge, access to data basis of genuine information which is classified according to its scientific value.

In today's information society, technological and digital training has become a necessity in order to prevent risks, information alienation and to ensure professional success and safe social coexistence.

2 Case Study Library of the "Lucian Blaga" University of Sibiu - Risks in Digitization Projects

From the general to the specific, in order to identify the risks faced by the library, we have conducted an evaluation, first by defining the library's mission, its general and specific objectives, the resources and activities used to fulfil its mission and to meet its objectives.

2.1 The Mission and the Objectives of the Library

The mission of the Library of the "Lucian Blaga" University of Sibiu is to purchase, organize and ensure access to a great variety of information, materials and services with a view to fulfilling the intellectual, information and research needs of all its users (students, professors, researchers) within our University and the local academic community.

Table 1: Objectives of the Library

<i>General objectives</i>	<i>Specific objectives</i>
Ensuring an adequate reference collection to support education and research programs	Purchasing publications fit for the LBUS educational offer, according to the Library's purchasing policy: every month, between 10 and 20 titles, depending on the books published and on the allocated funds
	Subscribing to printed internal and external periodic publications at the end of November for the following year (around 200 subscriptions)
	Expanding the LBUS digital collection by scanning and processing around 2,000 pages per year
	Participating to the Gaudeamus Book Fair

	each November
Creating and maintaining appropriate library facilities	Conducting in-library marketing studies at the end of each calendar year, by asking users to fill in questionnaires
Ensuring access to the library's collections and to its other services	Organizing weekly meetings with users, in groups of maximum 30 people, for specific information-retrieval services, guidance and access to using the data bases
Writing specific works to ensure that users are informed about existing documents, ways to retrieve information from and to consult catalogues, reference materials, reader conferences, thematic exhibitions, book presentations, newsletters	The quarterly publication of printed and online catalogues and reference materials regarding new entries
	Making available 100 reference lists to students every year, as per the recommendations of mandatory bibliography
Digitizing library-specific activities	Continuous management of the Library5 information system, tracking and ensuring that it is always operational
	Permanent usage of technology and of ITC in specific activities
Developing cooperation and document exchange relations with other libraries in Romania and abroad or with other academic, scientific and research institutions	Supporting internal and international exchange of LBUS professors' publications at the end of each quarter (around 100 copies to 60 partners)
Methodological and scientific support to the public libraries in the city and in the county of Sibiu	Quarterly meetings with 50 librarians in the education system in the county aimed at granting them methodological and scientific support
Ensuring the qualification, continuous development and recycling of its personnel through individual study, the organization of library science classes and sending them to the courses organized at central level by the competent ministry	Conducting training session for library employees, through annual meetings with members of the professional associations
Organizing conferences, scientific symposiums, experience exchanges on library science and bibliography-related topics, participation at local, national, and international specialized events, conducting	Organizing the annual International Conference on Information Science and Literacy

research in bibliography or in other domains of science and culture, collaborating with professional organizations of libraries in education and in other networks	
Concluding agreements, conventions, contracts regarding specific activities with Romanian and international bodies and organizations	Securing local, national or international projects every year, through direct competition or in partnership with various institutions.

It can be said that a library's mission is closely connected to its documentary collections, as it mostly focuses on their development, preservation and circulation. Thus, in order to identify the risks related to all its collections, the potential threats related to the information, research, documentation processes, as well as those that may prevent the fulfilment of these needs must be clearly understood.

Considering its mission and its objectives, the LBUS Library has defined the risks below:

Table 2: Risks in a Library.

<i>Risks</i>	<i>Risk-associated circumstances</i>
The risk of not providing continuous collection development	Library strategic planning / Operational plans for hardware
The risk of not making the collections available to the users in due and proper time	Current legislation
	Insufficient fund allocation / The changing priorities in the investment policy
The risk of physical degradation of the collections	Delays in publication / No longer editing a publication / Provider-related problems
	Availability of online information resources
The risk generated by the educational infrastructure	Changing priorities related to research / Delays in receiving the educational offer
Digital risk	Non-dissemination of information from higher structures to direct beneficiaries / Not understanding users' needs
	Insufficient / incompetent / uninterested personnel
	Faulty promotion

	Outdated technology / Equipment failure / Insufficient IT hardware / Improper management and maintenance of digital content
	Improper management of the physical space / insufficient infrastructure
	Storage conditions / Damage through use

Once the risks have been identified, the next step is to appoint the people responsible for the management of each risk and then to evaluate them based on the probability of their occurrence and their estimated impact. Each responsible will adopt a strategy and establish action plans and control instruments to mitigate each risk.

2.2 Digital risk

We shall now focus on **digital risks**, identifying and analyzing them in the context of the continuous changes which the library has undergone over the last two years, especially because of the new technologies implemented that have resulted in fully atomized activities and services.

Digital risk has been identified for the action domains below:

- preventing the discontinuation of the library activities: records, archiving, processing, statistics, circulation, and creation of digital content, intranet and e-mail access;
- the availability of digital information: online catalogue, subscribed data bases which can be accessed online, the LBUS Digital Library;
- digital data security: information content, digital content;
- protection of personal data: information about library personnel, users, copyright;
- protection against viruses, destructive software, cyber-attacks.

Digital risks are a result of threats by: people, through internal or external, intentional or unintentional actions; technology, directly connected to technical issues; or natural, independent from human action.

We have become extremely aware of digital risks and of disturbing factors for the activities we unfold following our participation at digitization projects, in which, besides the importance granted to human resources and to information technology, without which it would have been impossible to fulfil the objectives undertaken, we have identified other potential areas of risk:

- choosing the correct library management and digital library information system;

- choosing the proper server systems, both hardware and software;
- implementing an effective digitization project, choosing valuable documentation content that is suitable for the users and corresponds to the technical requirements, in line with the "Europeana" Digital Library;
- managing and taking part in activities and work meetings within the various projects;
- allocating an optimum amount of time to each activity;
- reallocating budget per activities;
- insufficient documentation regarding the work flow or method;
- incomplete documentation and technical requirements;
- unrealistic expectations;
- lack of previous experience;
- problems related to partners, collaborators, providers.

All these aspects had to be managed very efficiently: the risk evaluation was conducted starting from the planning phase, so as to allow the project manager to identify the weak or risky points and to plan strategies to deal with the actual occurrence of the risk.

Conclusions

In today's digitalized society, new information technologies have brought about new specific risks. In this paper, we have documented the general and the specific risks of the digital environment.

We then looked for, identified, listed, and analyzed the potential risks for the Library in Sibiu, which has undergone a modern development process through the implementation of modern library-specific information technologies.

On the one hand, we have the value of information and the importance of protecting it; on the other hand, there is the free movement and the accessibility of information on the internet. Our experience over the last two years of the digitalization process and projects has made us more aware of how to manage, process and keep information safe. New measures and information security systems are thus required, capable of excluding threats, regardless of whether they are man-made or of natural origin.

The use and the implementation of risk management helps institutions reduce the negative effects of risks and to achieve better results, exceeding initial estimates, thus supporting the organization to fulfil its objectives.

Risks cannot and should not be completely eliminated, they appear regardless of the experience accumulated, and when properly managed, can even be turned into opportunities.

In today's information society, technological and digital training has become a necessity in order to prevent risks, information alienation and to ensure professional success and safe social coexistence.

References

- [1] Burcea, Magdalena; Tanase Ion, *Controlul managerial și managementul riscurilor la entitățile publice (Managerial Control and Risk Management in Public Entities)*, Târgoviște: Valahia University Press, 2007
- [2] Dobrin, Gabriel Ionel, *Economia și evaluarea riscului în lumea afacerilor (Risk Economy and Evaluation in Business)*, "Lucian Blaga" University of Sibiu Publishing House, Sibiu, 2013, p.140
- [3] Ciocoiu, Carmen Nadia, *Managementul riscului, Vol. 1: Teorii, practici, metodologii (Risk Management, Vol. 1: Theories, Practices, Methodologies)*, Bucharest: ASE Publishing, 2008
- [4] Ciocoiu, Carmen Nadia, *Managementul riscului, Vol. 2: Modele economico-matematice instrumente și tehnici (Risk Management, Vol. 2: Economic and Mathematical Models, Instruments and Techniques)*, ASE Publishing House, Bucharest, 2007, p.189
- [5] "How's Life in the Digital Age? *Opportunities and Risks of the Digital Transformation for People's Well-being*", OCDE Study, February 2019
- [6] Saptarshi Ganguly, Holger Harreis, Ben Margolis, and Kayvaun Rowshankish - *Digital risk: Transforming risk management for the 2020s*; <https://www.mckinsey.com/business-functions/risk/our-insights/digital-risk-transforming-risk-management-for-the-2020s>
- [7] Bitten Thorgaard Sørensen - *Digitalisation: An Opportunity or a Risk?* Journal of European Competition Law & Practice, Volume 9, Issue 6, 1 June 2018, Pages 349–350, <https://doi.org/10.1093/jeclap/lpy038>
- [8] SR ISO/IEC 27005, *Information technology. Security techniques. Information security risk management*, Bucharest, ASRO Publishing House, 2016

Webliography

- [1] *Risk Management and Contingency Planning*, AHDS, <http://www.ahds.ac.uk/creating/information-papers/risk-management/>
- [2] Horava, Tony, *Risk Taking in Academic Libraries: The Implications of Prospect Theory* <https://journals.tdl.org/llm/index.php/llm/article/viewFile/7055/6280>
- [3] Michalko, James; Malpas, Constance; Arcolio, Arnold, *Research Libraries, Risk and Systemic Change*, OCLC Research, March 2010 www.oclc.org/research/publications/library/2010/2010-03.pdf

THE EUROPEANA COLLECTIONS – TRANSYLVANIAN PRINTS FROM THE COLLECTIONS OF THE NATIONAL MUSEUM OF THE UNION FROM ALBA IULIA

Ana Maria Roman Negoii¹,

*¹ "1 Decembrie 1918" University of Alba Iulia, Department of History,
Archeology and Museology, <http://diam.uab.ro>, Romania*

Abstract

The Europeana Collections, inaugurated in 2008, represents the completion of an ambitious project intended to be a journey and an instrument for the access to the culture, the history and the identity of the Europeans. Nowadays, Europeana contains in its collections more than 58 million digital units, organized on domains and themes, from art works, artefacts and books to movies and music. The patrimony of the Europeana enriches yearly and constantly by the contribution of the European Member States, as response to the common aspiration to open the access to knowledge beyond the national or territorial borders. The Europeana Project represents the implementation by digitization of a set of standards and of a unitary approach on the valorisation by digitization of the patrimony of the European states. The country reports reflect the most accurate the measures and the achievements of each contributing state to Europeana. Therefore, the Romanian report – a document updated in January 2019 – presents punctually the achievements of our country and of the Romanian institutions contributing to the enrichment of the Europeana collections. The list of the contributors contains, next to the names of well-known libraries, the name of the National Museum of the Union from Alba Iulia, with 986 digital units. Related to the field of the prints, the National Museum of the Union is not a direct contributor, but its collections are uploaded by other contributors. The National Museum of the Union has a remarkable and extremely valuable collection of Transylvanian books, mainly printed in Cluj during the 18th century. There are 78 titles with a preponderantly religious, juridical and educational content, representing an important segment of the national cultural heritage. The present paper aims to approach the above mentioned works and to identify them in the Europeana collections.

Keywords: Digital library, Europeana collections, Transylvanian prints

1. The Europeana Collections – the largest digital European project

Opened in 2008, Europeana has today over 58 million units in its collections, organised on domains and themes. And the number is growing day by day. The mission of Europeana is very simple: "We transform the world with

culture! We want to build on Europe's rich heritage and make it easier for people to use, whether for work, for learning or just for fun" [1].

Several milestones have been recorded in its evolution:

In January 2011, the European Commission released its New Renaissance report which endorsed Europeana as "the central reference report for Europe's online cultural heritage". In 2015, Europeana's collection-related websites e.g. its search engine, exhibitions and blog, started to come together in one place, now called Europeana Collections [2].

March 2017. The Expert Group on Digital Cultural Heritage and Europeana ("the Group") is set up. The Group's tasks: to review and discuss policies for digital cultural heritage, notably by assisting the Commission in monitoring and assessing the progress and the impact of the implementation of the European Commission Recommendation of 27 of October 2011 on the digitisation and online accessibility of cultural material and digital preservation (2011/711/EU) and related Council Conclusions, in particular those of 31 May 2016 on the role of Europeana for the digital access, visibility and use of European cultural heritage [3].

September 2018. In response to the Council's request, the Commission carried out an evaluation of Europeana based on the five mandatory criteria (relevance, effectiveness, efficiency, coherence, EU added value) of the Better Regulation Guidelines of the Commission.

Europeana has one of the largest digital collections in the world and is the only pan-European platform of its kind to provide access to image, text, sound, video and 3D material from the collections of over 3700 European libraries, archives, museums, galleries and audio-visual institutions. Europeana further stands out by offering material in a large number of languages (37 languages). Europeana's relevance to EU policies and priorities was rated high. Europeana fares well in some aspects but there is significant potential for improvement in others. The provision of common standards, best practices and promotion of open cultural data are considered among its key achievements. Europeana's common solutions and publishing frameworks have been taken up by cultural heritage institutions. However, the many user groups involved pointed out weaknesses in the technical infrastructure (including the aggregation infrastructure) and in the platform's functionality in terms of multilingualism search and filtering facilities in the portal and the APIs and the quality of the data.

1.1. Orientations for the future development of Europeana

The future scope of Europeana will be mainly focused on the needs of primary stakeholders in the cultural heritage sector, in particular the cultural heritage institutions, which the initiative should support and steer in their efforts towards achieving digital transformation. Europeana will strengthen its ties

with the cultural heritage institutions. Furthermore, the Commission considers that the Europeana initiative should be established around the empowerment of cultural heritage communities in Member States. Europeana should provide substantial support to data providers and aggregators for an efficient data ingestion and update. The Commission considers the aggregators —local, national and domain-related —to be essential to the Europeana initiative and considers, that their role as intermediaries between cultural institutions should be actively supported by Member States.

Europeana will continue to be funded under the CEF programme through a mix of grants and procurements until the end of the current multiannual financial framework. While the core services will be funded through procurements in order to assure the stability of the service provided, the financial instruments of grants will be used to support cultural institutions in order to contribute to Europeana. Under the next Multiannual Financial Framework (COM(2018) 321 final), the Commission has proposed to fund Europeana under the Digital Europe programme which will aim to strengthen the capacity building of various sectors and the flagship projects in these sectors making use of this capacity building. The work carried out under the this programme will be complemented through research and innovation activities in Horizon Europe on digitisation of cultural heritage monuments and sites, automatic interpretation of cultural heritage as well as virtual museums.

The evaluation represents a turning point, 10 years into Europeana's existence, with a number of achievements as well as shortcomings as described above. Europeana needs to renew its added value to the cultural heritage sector, for instance through capacity building, tools for high quality content and metadata supply and enrichment, an efficient infrastructure for data delivery and continuous support. To this end, the Commission will continue to support Europeana and its network of aggregators, institutions and professionals in their efforts to the digital transformation of the cultural heritage sector [4].

2. The National Museum of the Union in Europeana Collection

According to the Report of Romania on the 20th of January 2019, available on the site Europeana [5] the contribution of our country to Europeana is of 147.173 digital units and the contact point for Romania is National Heritage Institute.

The National Museum of the Union from Alba Iulia is among the Romanian contributing institutions to the Europeana project. There are only two Romanian museums contributing to Europeana, with the National Institute for Heritage as aggregator.

We depict in our paper the contribution of the National Museum of the Union from Alba Iulia, based on the following considerations: the importance of the

local and national heritage in the custody of the institution and its efforts to bring visibility to this heritage. Europeana represents visibility and multiple possibilities to connect with similar institutions in order to promote the cultural and historical heritage. Nowadays, the National Museum of the Union from Alba Iulia is present in Europeana with 986 digital units [6], representing religious artefacts, funerary artefacts and icons. Their aggregator is the National Heritage Institute.

National Museum of the Union from Alba Iulia also preserves in the Museum Library an old books collection fund of national importance. The Museum Library, established at the end of the 19th century, originally contained a fund of 1.000 volumes [7]. It contains today 67.126 volumes, organised on the following domains: Romanian and foreign periodicals, art, archaeology, history, literature, miscellaneous, ideology, philosophy, science, geography, ethnography, Romanian old books, foreign old books and the fund "The Old Library". A valuable fund of local and national periodicals (1922-2009) is also available in the Museum Library.

The funds of the library increased significantly in the period between the wars, while Ion Berciu was the manager of the institution. Starting from the initial funds of books and periodicals, three "special collections" formed: The Old Library, the Romanian old books collection – 684 units and the foreign old books collection – 349 units. We mention only a few titles that are significant for the history of prints and books in the Mediaeval and Modern period: *Pravila de la Govora*, 1640; *Evanghelia învățătoare*, Deal, 1644; *Biblia*, București, 1688; *Noul Testament*, Bălgrad, 1648; *Chiriadromionul*, 1699; *Pânea pruncilor*, 1702 [8].

The Old Library is important for the scientific field, possessing 78 titles of Transylvanian old books, more than half of them printed in Cluj in the 18th century. 23 of these titles have been classified since 2014 [9] as treasury items. The catalogue Transylvanian prints in the collections of the National Museum of the Union from Alba Iulia edited by Florin Bogdan and published in 2015, offer details on the 78 titles for those intending to find out more on the subject [10].

The 23 works classified as treasury items are related to religion, juridical studies, geography and history. They are relevant for the Transylvanian cultural heritage in the Age of Enlightenment and also for the Central European area. The classification of the books as treasury items confirms and increases their value. Still, legally, the classification generates a restriction of the access at the respective work regarding the research and display conditions in exhibitions. Digitisation proved to be the best solution in order to maintain the access to this kind of works. Thus, we followed the list of the 23 titles in Europeana. For the books enthusiasts, identifying a rare book or a limited edition of an old book represents an investigation similar to a detective work or to an archaeologist work. Therefore, due to the fact that Europeana grows day-by-day, enriching with the patrimony of European representative libraries, we investigated if there is possible to find copies or editions of the

books from the collection of the National Museum of the Union from Alba Iulia in the large digital collection of Europeana. From the list of 23 treasury books we identified 3 titles in Europeana:

2.1. Erdély országának Három könyvekre osztatott törvényes könyve. Melly aprobata, compilata constitutiokból, és novellaris articulusokból áll

The work contains the well-known law corpus that regulated the juridical situation of Transylvania in the 16th-18th centuries: *Approbatae Constitutiones* and *Compilatae Constitutiones*. *Approbatae Constitutiones* is a synthesis of the decisions of the Transylvanian Diet during the Principality, 1540-1653, and represents the most important document of Transylvanian law for the 17th century[11], while *Compilatae Constitutiones* covers the period 1654-1669. These collections of laws remained in force after the Austrian military occupation and received an official recognition from the Habsburg power through *Diploma Leopoldina* (1691). The Transylvanian privileged classes used these legal provisions, especially during the 18th century, in order to defend the rights they had previously gained, in the relation with the imperial authorities[12]. Their originality consists in the fact that they are not a translation of foreign laws, but a local juridical construction, with particularities specific to Transylvania.

The National Museum of the Union from Alba Iulia possesses two copies of the 1779 edition[13], printed to the Typography of the Reformed College of Cluj, and also two copies of the edition from 1815-1816 [14], printed to the Typography of the Royal College of Cluj. The edition from 1815-1816 was proposed in 2016 for classification in the category of the cultural mobile goods.

The edition from 1779 is formed from: Part I, *Aprobates Constitutiones Regni Transilvaniae & Partium Hungariae eidem annexarum*, Part II, *Compilatae Constitutiones Regni Transilvaniae & Partium Hungariae eidem annexarum*, and Part III, *Novellarum articulosok* (New articles). Last part comprise decisions of the Transylvanian Diet between 1744-1755.

The edition from 1815-1816, printed in Cluj, appears under the title *Erdély országnak három könyvekre osztatott törvényes könyve, melly aprobata, compilata constitutiokból és novellaris articulusokból áll* and is formed from: "*Novellaris articulosok*", "*Index novellarium articulorum, diaetalium ab anno 1744 usque annum 1792*", Cluj, 1816; "*Statuta jurium municipalium Saxonum in Transylvania*", Cluj, 1815, and "*Index Statutorum seu Jurium Municipalium Saxonum in Transylvania*", Cluj, 1816.

We can find in the Europeana collections the digital format of the integral 1815-1816 edition from Cluj, with The European Library as aggregator, available to the Bavarian State Library [15]. The book contains *Novellaris articulosok*, *Aprobata* and *Compilata* (Table of contents), *Index Novellarium articulorum, diaetalium ab anno 1744. usque annum 1792*, *Approbatae*

Constitutiones and Compilatae Constitutiones. Halmágyi Istvan, secretary of the Transylvanian Gubernium, appears as author of Index Novellarium articularum.

Understanding the Medieval Transylvanian law and its evolution to modernity is an important subject, offering interesting directions of research. The integral availability of the primary sources, in digital format, encourages the research, in total agreement with the motto imposed by the Enlightenment: Sapere aude – Dare to know.

2.2. Instructio pro tabula regia Judiciaria Transylvanica

The second book is also a juridical work, printed in 1777 in Cluj-Sibiu, at Hochmeister Typography, as indicated by the title page. The reference to the printing place can be explained by the fact that the Hochmeister family owned at the end of the 18th century – the beginning of the 19th century more libraries in Transylvania, at Sibiu, Cluj or Blaj [16]. The National Museum of the Union from Alba Iulia possesses one copy [17]. The book is available integrally in digital format to the Romanian National Library and to the Austrian National Library. The aggregators are the Romanian National Library and, respectively, The European Library [18].

Instructio pro tabula regia Judiciaria Transylvania presents an Instruction for the organisation of the most important law court from Transylvania.

Tabula regia (The Royal Board) was the Supreme Court in Transylvania. It functioned from 1754 in Târgu Mureş, previously functioning in Mediaş [19]. The Instruction was issued by Empress Maria Theresa and approved by the Transylvanian Diet on the 6th of August 1777 [20]. It refers to the organisation and the functioning of the court. Documented for the first time in 1542, The Royal Board functioned until 1890, when it was abolished by the Law XXV related to the Appeal Courts [21]. There was a 4 year period between 1786-1790 when Tabula regia was moved to Sibiu by the emperor Joseph II [22]. The most important court in Transylvania, the Royal Board had the jurisdiction to judge inclusively the treason offences. It also judged the appeals against the solutions of the County Courts, the Transylvanian Saxon University and the Szeckler Courts. When the solution was against the solution of the inferior courts, the parties can appeal to Gubernium, and, exceptionally, after the Gubernium, to the Aulic Chancellery or even to the Emperor, who was the last decisional instance.

The book is a very important study instrument for understanding the evolution of the judicial system in Transylvania in the 18th century and for the amplitude of the Habsburg reformism [23].

2.3. Concordia orthodoxorum patrum orientalium et occidentalium De Spiritus Sancti Processione, ex

**Comentarijs Gennadij Patriarchae Constantinopolit:
excerpta per reverendissimum dominum
Christophorum Peichich Missionarium Apostolicum,
nec non Abbatem S. Georgii de Csanád, Laureatis
Honoribus Specabilium Praenobilium, Reverendorum
Nobilium, ac Eruditorum DD. Neo-Baccalaurerorum,
Dum In Alma, ac Regio Principali Universitate S. J.
Claudiopolitana promotore R. P. Michaelae Salbeck e
S. J. AA. LL & Philosophiae Doctore, ejudémque
Professore Ordinario Prima AA. LL. & Philopsophiae
Laurea insignirentur a condiscipulis dicata Anno M.
DCC. XLV [1745].**

The third title owned by The National Museum of the Union from Alba Iulia and identified in Europeana is Concordia Orthodoxorum Patrum Orientalium et Occidentalium, printed at the Typography of Jesuit Academic Society from Cluj in 1745 [24]. It has 112 pages. We find the entire work in digital format at three locations: The Romanian National Library, with the Romanian National Library as aggregator, the Bavarian State Library and the Austrian National Library, with The European Library as aggregator [25].

The work was printed for the first time in Trnava (Slovakia) in 1730 [26]. The second edition appeared in 1745. The work has a theological content and supports the policy for the consolidation of the Catholicism by the Habsburg Empire. The author is Christophorus Peichich, an Apostolic missionary, Bulgarian by his origin.

Peichich was the author of four books, divided in two groups. The first group concerns the question of the schism between Eastern and Western Churches and consists of three publications: Zarcalo istine (The Mirror of Truth between the Eastern and Western Churches), Speculum veritatis (The Mirror of the Truth) and Concordia orthodoxorum Patrum orientalium et occidentalium. Zarcalo istine (Venice 1716) was Peichich's first publication and was written in a variant of Southern Slavic ("Illyrian") language. Speculum veritatis (Venice 1725) was an enlarged version of the Mirror in Latin. Concordia orthodoxorum Patrum orientalium et occidentalium (Trnava 1730) was a closer examination of an aspect of the question. All three are designed according to the literary genre of controversistic theology, yet seem animated by a concordistic spirit, aiming to promote the reunification of the Eastern and Western Churches. The second "group" of works actually consists in only one publication: the Mahometanus in lege Christi, Alcorano suffragante, instructus (Trnava 1717), a "catechism" for Catholic missionaries carrying out their activity among Muslims.

Within the context of the Hapsburg's policy, a policy aimed at the religious integration of their subjects and the consolidation of Catholicism as the "state religion" in their Empire, Peichich's books had to serve the cause of the union

of the "schismatic" Orthodox Church with the Catholic Church and of converting Lutheran and Calvinist "heretics" and Muslim "infidels" to Catholicism.

According to Peichich's view, the European Powers were to join forces under the leadership of the Catholic emperor in order to form a Christian front united against the Ottoman Empire. The project of the author as a missionary and a man of letters was to contribute to the fulfillment of that end by providing the militia christiana with the "spiritual weapon" of his polemical works [26].

Conclusions

We started our investigation encouraged by the Europeana mission: "we aim to transform the world through culture". The digital technology of the 21st century and the effort of the libraries, museums and cultural institutions to capitalize and make visible the cultural and historical inheritance transform Europeana in the largest and most dynamic cultural resource at the present time. It creates working instruments and connections beyond the geographical borders for all researchers and, in general, for all people.

References

For the references' section, follow the next example for book and article:

- [1] <https://www.europeana.eu/portal/en> Consulted on 06.05.2019.
- [2] <https://pro.europeana.eu/our-mission/history> Consulted on 06.05.2019.
- [3] http://ec.europa.eu/information_society/newsroom/image/document/2017-42/commission_decision_dche_D19B28A2-BCEE-B2D6-81F1AA9FB3CE377C_47767.pdf Consulted on 06.05.2019.
- [4] <https://eur-lex.europa.eu/legal-content/RO/TXT/?uri=CELEX:52018DC0612;>
- [5] <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0612&from=EN> Consulted on 06.05.2019.
- [6] Country-Report-Romania-January-2019 <https://pro.europeana.eu/what-we-do/member-states> Consulted on 06.05.2019.
- [7] https://www.europeana.eu/portal/en/search?q=Alba+Iulia&f%5BDATA_PROVIDER%5D%5B%5D=Muzeul+Na%C8%9Bional+al+Unirii+-+ALBA+IULIA Consulted on 06.05.2019.
- [8] Valer Moga, Eva Mârza, "Muzeul Unirii din Alba Iulia", in *Apulum*, XXVII-XXX/1990-1993, p. 421.
- [9] Carmen Stănea, Biblioteca Muzeului Național al Unirii din Alba Iulia, <http://www.bibnat.ro/biblioteci.php?id=3358> Consulted on 06.05.2019.

- [10] National Heritage Institute, <http://clasate.cimec.ro/lista.asp?det=8213-Muzeul-National-al-Unirii-ALBA-IULIA>; http://clasate.cimec.ro/omc/OMC-2822_28-11-2014.PDF; http://clasate.cimec.ro/omc/OMC-2816_28-11-2014.PDF Consulted on 07.05.2019.
- [11] Florin Bogdan, *Tipărituri transilvănene din colecțiile Muzeului Național al Unirii Alba Iulia*, Editura Altip, Alba Iulia, 2015.
- [12] Cătălin Bichescu, *Proceduri judiciare și administrative în Transilvania (secolul XVIII)*, Cluj-Napoca, Editura Mega, 2017, p. 26. For *Approbatae*, there is available in Romanian the translation of the original texts, with important historical and juridical. See Alexandru Herlea, Valeriu Șotropa, Romul Pop, Iuliu Nasta, Ioan. N. Floca, *Constituțiile Aprobate ale Transilvaniei (1653)*, Cluj-Napoca, Editura Dacia, 1997.
- [13] *Ibidem*, p. 25-26.
- [14] Quotations for the two copies of the 1779 edition are: BV 1062 coll.1 and I 895. The 1779 edition is bound with *Statuta Jurium Municipalium Saxonum in Transylvania*, Cluj, 1779. See Florin Bogdan, *Tipărituri transilvănene din colecțiile Muzeului Național al Unirii Alba Iulia*, Editura Altip, Alba Iulia, 2015, p. 42-43.
- [15] Quotations for the two copies of the 1815-1816 edition are: CVR 493 (volume 2) and CVS 85 (volume 3-4).
- [16] <https://www.europeana.eu/portal/en/search?q=Erd%C3%A9ly+orsz%C3%A1g%C3%A1nak+H%C3%A1rom+k%C3%B6nyvek+osztatott+t%C3%B6rv%C3%A9nyes+k%C3%B6nyve> Consulted on 07.05.2019.
- [17] Florin Bogdan, Simona Loredana Bogdan, "Valori bibliofile în fondul documentar al Bibliotecii Municipale „Petru Maior” Reghin", in *Libraria*, Anuar X, 2011, p. 4, <http://www.diacronia.ro/ro/indexing/details/A19462/pdf> Consulted on 13.05.2019.
- [18] Florin Bogdan, *Tipărituri transilvănene din colecțiile Muzeului Național al Unirii Alba Iulia*, Editura Altip, Alba Iulia, 2015, p. 41.
- [19] <https://www.europeana.eu/portal/en/search?q=Instructio+pro+tabula+regia+Judiciaria+Transylvanica> Consulted on 13.05.2019.
- [20] For the evolution of the courts in 1542-2004 in Transylvania see <http://portal.just.ro/43/SitePages/prezentare.aspx#Istoric>
- [21] Cătălin Bichescu, *Proceduri judiciare și administrative în Transilvania (secolul XVIII)*, Cluj-Napoca, Editura Mega, 2017, p. 21.
- [22] *Ibidem*, p. 17.
- [23] Liviu Moldovan, "Tabla regească din Transilvania", in *Revista Arhivelor*, anul LII, vol. XXXVII, nr. 2, București, 1975, p. 197-201.
- [24] For the organization of the judicial life in Transylvania, see Nicolae Balint, *Aspects regarding the organization and the function of the Transylvanian*

Appeal Court of Târgu Mureș,
https://old.upm.ro/facultati_departamente/ea/RePEc/curentul_juridic/rcj08/recjurid081_212F.pdf

- [25] The National Museum of the Union from Alba Iulia possesses two copies, quotation BV 378 and CVS 174. See Florin Bogdan, *Tipărituri transilvănene din colecțiile Muzeului Național al Unirii Alba Iulia*, Editura Altip, Alba Iulia, 2015, p. 30-31.
- [26] <https://www.europeana.eu/portal/en/search?q=Concordia+orthodoxorum+patrum+orientalium+et+occidentalium>+ Consulten on 13.05.2019.
- [27] Christophorus Peichich, *Concordia orthodoxorum Patrum orientalium et occidentalium in eadem veritate, de Spiritus Sancti processione ab utroque, ad amussim convenientium: ex commentariis Gennadii Patriarchae Constantinopolitani excerpta...*, Tyrnaviae: Typis Academicis per Fridericum Gall, 1730. Related to the author of the book, Christophorus Peichich, an Apostolic missionary, see Iva Manova, *The Cultural Project of Krastyo Peykich (1666-1730): A 'Spiritual Weapon' for the Catholic Undertaking in Eighteenth-Century East Central Europe*, PhD Thesis, p. 22, http://paduaresearch.cab.unipd.it/5153/1/Tesi_Manova_PDF-A.pdf Consulted on 13.05.2019.

SUGGEST RECOMMENDATION FOR LIBRARY USERS USING GRAPHS

Gheorghe-Cătălin Crișan¹,

¹ PhD. student, University „Lucian Blaga” of Sibiu, Faculty of Science, Romania

Abstract

The aim of this paper is to prove the usefulness of graphs in solving an ever-present problem for library users: finding books they like and they are looking for. Graphs are known as an important tool in solving conditioned optimization problems. We propose a graph-based system of recommendation which can be easily used in a library for assisting and helping users in finding in real time the books they like. The main advantage of the proposed graph-based approach lies in the ease with which new data or even new entities from different sources are added to the graph without disturbing the entire system. The system uses the similarity scores in order to find the similarity between objects and to get the best recommendation for a user's request. In the end, we will compare the results from used formulas..

Keywords: graph, optimization, similarity, library

1 Introduction

Graphs are everywhere. They are known as an important tool in solving conditioned optimization problems. For example, Google Maps use graphs to find the best routes for cars, buses, and walks. Thus, we need to find a different approach for each type of route based on some requirements. Car drivers have to respect some traffic rules like speed limit, one-way streets, crowd and so on. All of these influence the time of arrival to destination. For buses, we have to know the routes of these, when the buses come in a specific station, even the time spent by the user during the route. This is useful when the user has to change two buses to reach the destination and we want the user to spend as little time as possible. Also for the walks, we have to know about restricted walk area or even the weather, so the user can bypass a rainy area.

Search engines use graphs to rank web pages. This feature allows us to get only relevant pages based on searched keywords. So, if a user searches for pet stores we want to show only pages for pet stores near the user location for example.

Another good example is social media networks. We have a graph which looks like a network where the users are connected based on some rules. If a user is friend with other user the graph will create an edge between them to make a

correlation. This is useful when we want to recommend new friends or possibly known peoples.

Nowadays, a large part of the commerce takes place in the online environment, which is why we want to find the most effective ways to increase sales. A very important factor is providing personalized recommendations to every user at the right time or even in real time. This ensures that the suggested products have a high level of interest for the consumer, so he is willing to purchase the product, and the business can achieve an increase in the percentage of sales of the products.

Product recommendation is not just a sales strategy, it's also an action that helps improve the user experience in the online environment. Thus, a pleasant experience can increase the number of conversions and sales and increase the potential ROI of marketing efforts to minimize the effort the user is making. For example, 35% of Amazon's revenue is generated by its recommendation engine [1].

The use of graphs for recommendations presents a number of advantages such as the ability to suggest real-time recommendations based on the latest user actions, ease of setting parameters to be taken into account when suggesting a recommendation, the ability to add data in graph from different sources (relational or nonrelated database, csv files, etc.) without compromising the already existing graph, but also the ease of integration with the existing application systems.

After this, we will go through formulas used for our experiment and we'll make a comparison between them.

The rest of the article is organized as follows: the second section presents the problem definition and why such a problem exists. The third section presents the proposed solution and what are the main benefits. In the fourth section we take a look at theoretical aspects of graphs and how we use the graphs to map the data and user data sets. The fifth section contains information about the tool used to map data and view them using Cypher and details about the results obtained using different mathematical approaches. And in the last section the conclusions and further research directions.

2 Problem definition

The main problems with libraries are that there is not so much interaction and also the reader only reads books recommended by friends.

Many libraries do not interact enough with readers which purchase books to know their preferences. Thus, the library is just a place where you come if the reader needs a book which he already knows the title. Only a small part of the readers are willing to read from other genres because there is a big chance they will do not like it. They will not risk with these books because they don't want to spend time for something that they will not enjoy. So, the readers

need somebody to know what they like to read and what they would like to read.

The second fact is that most of the readers tend to read books that are recommended by friends or written by the same author whose books they have read before and liked. This is because they know their friends preferences, they talk together about the books they read. Based on that, they will like to receive suggestions from them.

3 Proposed solution

The solution proposed in this article is a graph-based system of recommendation which can be easily used in a library for assisting and helping users in finding in real time the books they like.

The main advantages of this solution are:

- increased user experience because they receive good recommendation
- easy way to add new data to graph system from different sources like dumped database, .csv file or any other file with data
- easy to implement as an external micro service for recommendation

4 Theoretical Aspects

4.1 Applying Graph Theory

The graph is a mathematical structure with countless applications in real life. Based on Jonathan L. Gross et al. (2004) we will describe the main ideas of graph theory. Using the graph we can create relationships between different objects using the following elements of the graph:

- nodes: representing the objects that make up the data set
- edges: representing the relationship between two objects that have a certain connection

The nodes or entities we are going to use can be of several types. Thus, for example, we will have user nodes and book nodes, all interconnected according to the established relationships. In this way, if an X person buys a Y card, then an edge is created from node X to node Y.

In our case, the edges are to be unidirectional so that if node A has a relation (edge) with node B it does not mean that node B has a relation with node A. If we want to do this we use two edges, one from node A at node B and one from node B to node A, as can be seen in Fig. 1.

Another essential aspect is that each of the nodes represents an object that has a set of properties specific to each object type. Thus, a book object can have properties such as title, author, genre and publishing house. These sets of properties also apply to edges between nodes. So, each edge represents a

specific type of relationship that has a set of properties. For example, a "BOUGHT" type edge may have properties such as the acquisition date or rating offered for that product.

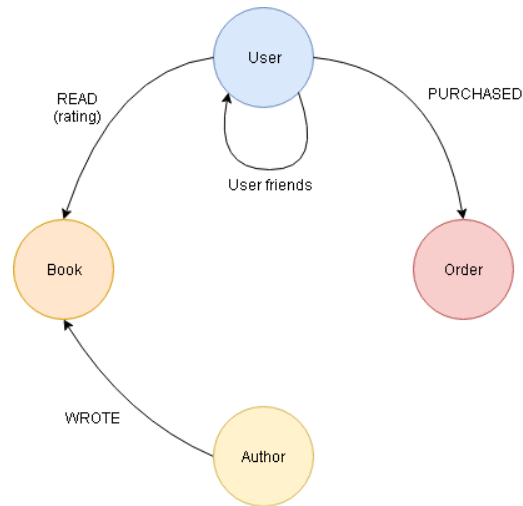


Fig. 1 – Example of graph with 3 types of nodes and 3 types of edges [2]

Thus, for the recommendation of books to the users of a library we have to have the following structure:

- the graph nodes:
 - o book (with properties):
 - title
 - authorId
 - genreId
 - o author
 - name
 - o genre
 - name
 - o user
 - name
 - years
 - gender
 - job

- the edges of the graph with various actions:
 - o wrote
 - o in_genre
 - o bought
 - rating

4.2 Types of Recommendations and Formulas

Suggestion recommendations can be of two types: content-based or collaborative filtering [3]. Content-based filtering also has features comparing objects properties and making a score based on them, where each feature can have a weight more or less important than another feature. Using this approach, the object with the highest score will be recommended, considering the user's preferences remain constant. The problem with this approach is the fact that suggestions are suggested only from the categories that the user has bought without recommendations for products in other categories, so that the user may be interested in them but have not seen them before.

On the other hand, collaborative filtering is based on how other users responded to the same object as compared to the current user. This determines whether our client might like a particular product (Guy, N. N., 2017). This is done by filtering users who have interacted with the same objects and finding similar objects that have been purchased by other users with similar preferences by doing a filter after the best score of objects.

The following formulas are also described in Junmei Feng et al. (2018) article presenting different algorithms for finding similarity. Using the Jaccard index we can measure the level of similarity between two objects resulting in a score with the value in the range [0, 1]. This means that two identical objects have the score 1 and two different objects altogether have a score of 0. The Jaccard index counts the common properties of two objects, that is, the intersection of the two sets of properties, then divides them into the total number of unique object properties, that is, the meeting of the two sets of properties. This can be applied for content-based recommendations.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Another function is the cosine distance with which we can compare the level of similarity between two objects resulting in a score that has a value in the range [-1, 1]. It transforms the values of the relations of the two objects into two vectors, and then calculates the distance representing the difference between the two objects. Thus for a 0° difference we have $\cos(0^\circ)=1$ meaning that the two objects are perfectly similar and $\cos(180^\circ)= -1$ meaning the two objects are totally opposite. It can be applied for collaborative filtering recommendations.

$$C(A, B) = \frac{\sum_{i=1}^n A_i * B_i}{\sqrt{\sum_{i=1}^n A_i^2} * \sqrt{\sum_{i=1}^n B_i^2}} \quad (2)$$

Pearson correlation is another feature used for collaborative filtering that takes into account the fact that for each object the value of relationships can differ, for example, two people can give a different rating for a book because one person is more demanding. Therefore, Pearson correlation takes into account

the average of values, so some objects tend to have higher values of relationships than other objects.

$$P(A, B) = \frac{\sum_{i=1}^n (A_i - \bar{A}) * (B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2 * (B_i - \bar{B})^2}} \quad (3)$$

5 Experimental setup

We have 3 main steps used in our project implementation to create the recommendation system. The first step is to create the dataset. Because we can't find an existing dataset for a library with users, borrowed books, ratings and so on we create a dataset with imaginary data. We put all these data in separated .csv file (one file for each entity – like books, author, reader, order).

The second step is to import created data into a graph database. There is plenty of graph database alternative: Neo4j, Titan, Cassandra, OrientDB, etc. before.

The third step is to choose a query-programming language. This depends on what graph database you choose: Cypher Query Language for Neo4j, Gremlin for Titan, Cassandra Query Language for Cassandra. So, for our project, we will use Cypher and we have a syntax like the one presented below:

```
(user: User {name: "Alex"}) -> [b:BORROW] -> (book:Book)
```

Fig. 2 – Listing example for Cypher query matching all books borrow by user Alex

Cypher Query Language is a declarative graph query language that allows for expressive and efficient querying and updating of a property graph [4]. The Cypher type system is simple to use and use a specific syntax for declaring nodes, relationships, paths, maps, lists, integers, floating-point numbers, booleans, and strings:

- CREATE / DELETE: Used to create / delete nodes and relationships.
- SET / REMOVE: Used to set / remove values to properties on nodes and relationships.
- MERGE: Used to match existing or create new nodes.
- MATCH: Used to get data from the graph
- WHERE: Used to add for filter results
- WITH: Used to passing results or to give aliases to results
- RETURN: Used to get results

Data sets are to be loaded from .csv files as you can see in the example below in Fig.3.


```

LOAD CSV WITH HEADERS
FROM 'file:///books.csv'
AS line
MERGE (book:Book { id: line.id,
                  title: line.title,
                  authorId: line.authorId,
                  genreId: line.genreId
                })

```

Fig. 3 – Loading book data set

6 Experiments and results

Based on data imported using the technique exemplified above we will create a graph like the next one. Here we can see authors and the books borrowed/bought by Maria Anders and Thomas Hardy represented as nodes:

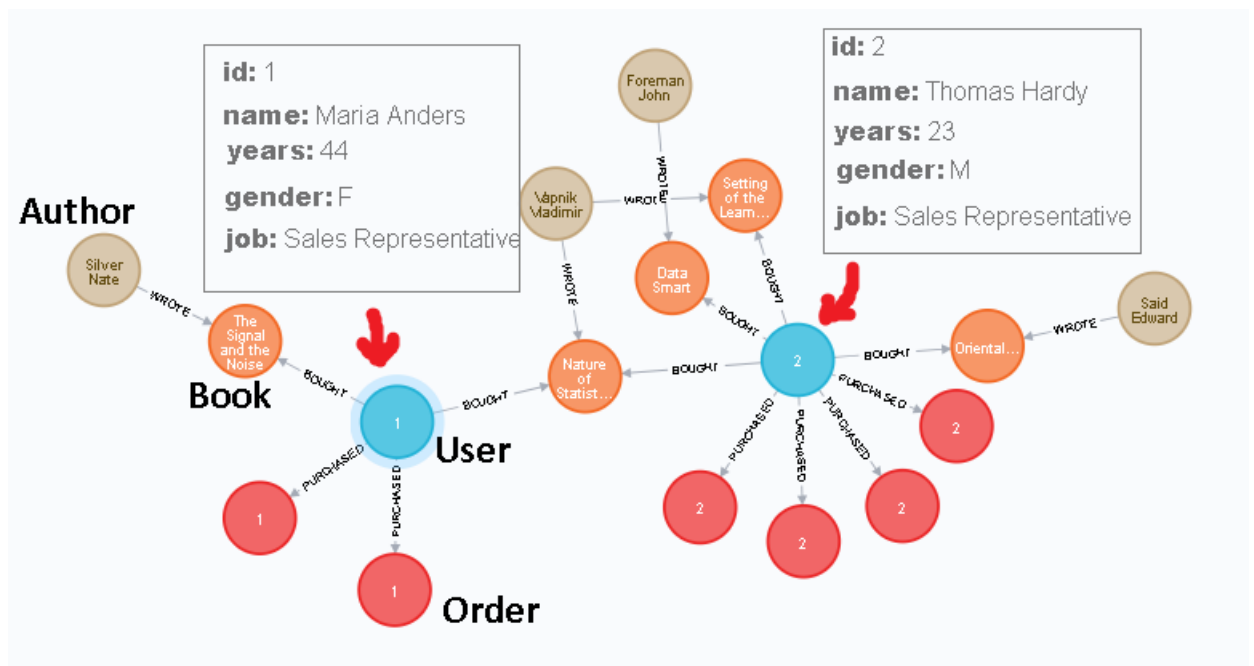


Fig. 4 – View graph relationships between two users

In the following section, we will describe formulas used for recommendations and what are the advantages and disadvantages of each one.

6.1 Jaccard Index

In the following, the recommended recommendations for the book "Nature of Statistical Learning Theory" using the Jaccard Index will be analyzed. The algorithm can be configured to take account of one or more features of objects. Thus, as can be seen in Fig. 5 the algorithm searches for similar

books that are the same or that are written by the same author. So, the higher the number of features in common, the better the score. However, this type of content-based recommendation is not very effective because it does not take into account which books other users are interested in, but just try to find cards with identical features.

The results obtained in Fig. 6 shows that the "Setting of the Learning Problem" book has the best score. This is due to the fact that the two books have the same author and the same genre, while the following recommendations have the same genre but different authors.

```
MATCH (b: Book {title: "Nature of Statistical Learning Theory"})-
[:IN_GENRE|:WROTE]-(g)<-[:IN_GENRE|:WROTE]-(other:Book)
WITH b, other, COUNT(g) AS intersection
MATCH (b)-[:IN_GENRE|:WROTE]-(bg)
WITH other, intersection, COLLECT(bg.name) as set1
MATCH (other)-[:IN_GENRE|:WROTE]-(og)
WITH other, intersection, set1, COLLECT(og.name) as set2
WITH other, intersection, set1, set2
WITH other, intersection, set1+filter(x IN set2 WHERE NOT x IN set1)
AS union
RETURN          other.title          AS          recommendation,
((1.0*intersection)/SIZE(union)) AS score ORDER BY score DESC LIMIT
5
```

Fig. 5 – Using of the Jaccard Index algorithm

recommendation	score
"Setting of the Learning Problem"	0.5
"Machine Learning for Hackers"	0.3333333333333333
"Analysis - Vol I"	0.3333333333333333
"Introduction to Algorithms"	0.3333333333333333
"Data Smart"	0.3333333333333333
"Fundamentals of Wavelets"	0.3333333333333333
"Python for Data Analysis"	0.3333333333333333

Fig. 6 – The recommendations obtained by applying the Jaccard Index algorithm

The advantage is that this approach is good when we want to match as many book properties as we can. So, we are looking for common properties like author, the genre of the book, etc.

The disadvantage is that there is no possibility to suggest books from other genres or even other authors. All these because the compared books will have a low score of similarity, so the recommendation will fail.

6.2 Cosine Distance

In Fig. 7 you can see the Cosine Distance algorithm listing on our data set. The algorithm suggests recommendations for the "Maria Anders" user looking

for other users who bought books that Maria bought. Thus, the algorithm performs collaborative filtering taking into account the preferences of other similar users. This is deduced from the user rating on shared cards. In this way, a classification of users who gave similar recharges for the books that the two users bought.

So, "Maria Anders" bought the book "Nature of Statistical Learning Theory" (rating 5) and the book "The Signal and the Noise" (rating 3). From the results obtained in Fig. 8 we can see that the algorithm suggests Mary as the first recommendation the book "Complete Sherlock Holmes - Vol I" bought by "Christina Berglund" (rating 5) who also bought the book "Nature of Statistical Learning Theory" (rating 3). This pattern shows that the two people have a book in common, and because the rating values given for the other books are similar, the algorithm finds this match.

```

MATCH (u1:User {name: "Maria Anders"})-[x:BOUGHT]->(b:Book)<-
[y:BOUGHT]-(u2:User)
WITH SUM(x.rating * y.rating) AS ab_sum,
      SQRT(REDUCE(ai = 0.0, a IN COLLECT(x.rating) | ai + a^2)) AS
a_sqrt,
      SQRT(REDUCE(bi = 0.0, b IN COLLECT(y.rating) | bi + b^2)) AS
b_sqrt,
      u1, u2
WITH u1, u2, ab_sum / (a_sqrt * b_sqrt) AS cos_distance
ORDER BY cos_distance DESC LIMIT 10

MATCH (u2)-[r:BOUGHT]->(m:Book) WHERE NOT EXISTS ((u1)-[:BOUGHT]-
>(m))
RETURN m.title AS recommendation, SUM(cos_distance * r.rating) AS
score
ORDER BY score DESC LIMIT 5

```

Fig. 7 - Using of the Cosine Distance algorithm

recommendation	score
"Complete Sherlock Holmes - Vol I"	5.0
"Making Software"	5.0
"Machine Learning for Hackers"	5.0
"Data Smart"	4.0
"The Trial"	4.0
"Python for Data Analysis"	4.0

Fig. 8 – The recommendations obtained by applying the Cosine Distance algorithm

The advantage is that this approach is to match users with similar properties such as books borrowed by both having the same rating or users have the same age and so on. So, now we can find a book written by other authors and even from other genres because a similar user read that book and now we can recommend to our reader.

The disadvantage is that this is not so accurate when we have some readers who like a niche book genre and we do not have other readers with similar preferences. Thus, for our niche user, we can't find other users with the same

preferences and we have to use the Jaccard Index instead to find similar books.

6.3 Pearson Correlation

The following implementation shown in Fig. 9 of Pearson Correlation comes as an optimization of the Cosine Distance algorithm. This is because the algorithm takes into account an average of the user's ratings, so the algorithm takes into account the overall direction in which the rating values tend. So the algorithm solves the recommendation problem by finding similar books that two people buy, even if a person is more exigent on the rating.

This is shown in Fig. 10 where "Maria Anders" bought the book "Nature of Statistical Learning Theory" (rating 5) and the book "The Signal and the Noise" (rating 3), with an average rating of 4. And the "Francisco Chang" bought the book "Nature of Statistical Learning Theory" (rating 4), "Machine Learning for Hackers" (rating 5), "Superfreakonomics" (rating 3) and "Physics & Philosophy" (rating 3) with an average rating of 3.75. Thus, applying Pearson's formula results in a coefficient equal to 1 which means that there is an absolutely positive linear correlation. This leads to the recommendation of a book read by "Francisco Chang", which is chosen from the books that "Maria Anders" did not read and whose rating of "Francisco Chang" is the highest.

```
MATCH (u1:User {name:"Maria Anders"})-[r:BOUGHT]->(m:Book)
WITH u1, avg(r.rating) AS u1_mean
MATCH (u1)-[r1:BOUGHT]->(m:Book)<-[r2:BOUGHT]-(u2)
WITH u1, u1_mean, u2, COLLECT({r1: r1, r2: r2}) AS ratings MATCH
(u2)-[r:BOUGHT]->(m:Book)
WITH u1, u1_mean, u2, avg(r.rating) AS u2_mean, ratings UNWIND
ratings AS r
WITH sum( (r.r1.rating - u1_mean) * (r.r2.rating - u2_mean) ) AS
nom,sqrt( sum( (r.r1.rating - u1_mean)^2) * sum( (r.r2.rating -
u2_mean) ^2)) AS denom, u1, u2 WHERE denom <> 0
WITH u1, u2, nom/denom AS pearson
ORDER BY pearson DESC LIMIT 10
MATCH (u2)-[r:BOUGHT]->(m:Book) WHERE NOT EXISTS( (u1)-[:BOUGHT]-
>(m) )
RETURN m.title AS recommendation, SUM( pearson * r.rating) AS score
ORDER BY score DESC LIMIT 5
```

Fig. 9 - Using of the Pearson Correlation algorithm

recommendation	score
"Machine Learning for Hackers"	5.0
"Orientalism"	4.0
"Data Smart"	4.0
"Setting of the Learning Problem"	3.0
"Physics & Philosophy"	3.0
"Superfreakonomics"	3.0

Fig. 10 – The recommendations obtained by applying the Pearson Correlation algorithm

The advantage is that this approach is good to match readers with common properties such as books borrowed by both with the same rating. Another advantage is also that we take into account the mean of the ratings for example. In this way, some users tend to give higher ratings than others so we have to find the correct pattern.

The disadvantage is the same as for cosine distance.

Conclusions

This article aims to present how to use graphs to suggest book recommendations to library users. It proposes an approach where the library application uses a graph saving data structure to attract as many users as possible.

On my opinion, the proposed application could be an important factor in loyalty to users because they receive accurate, real-time recommendations without the need for additional resources. Thus, the article has been able to show the benefits of using different statistical algorithms to find the level of similarity between two objects and how to implement them in real life to increase the number of hits of the library application.

One useful extension consists of using a hybrid recommendation system that uses the algorithms shown above but which also takes into account the latest user-viewed products as well as the number of times the product was viewed.

References

- [1] Sales force, *Product Recommendation Engines to Improve Customer Relationships*, <https://www.salesforce.com/solutions/industries/retail/resources/product-recommendation-engines> , (accessed 16 March 2019).
- [2] Microsoft, *Create a graph database and run some pattern matching queries using T-SQL*, <https://docs.microsoft.com/en-us/sql/relational-databases/graphs/sql-graph-sample?view=sql-server-2017> , (accessed 16 March 2019)..
- [3] Valiance Solutions, *RECOMMENDER SYSTEMS 101*, <https://valiancesolutions.com/recommender-systems-101> , (accessed 16 March 2019).
- [4] Wikipedia, *Cypher Query Language*, https://en.wikipedia.org/wiki/Cypher_Query_Language , (accessed 16 March 2019).
- [5] Jonathan L. Gross, Jay Yellen (2004). *Handbook of Graph Theory*, CRC Press
- [6] Junmei Feng , Xiaoyi Fengs, Ning Zhang, Jinye Peng (2018). *An improved collaborative filtering method based on similarity*, <https://doi.org/10.1371/journal.pone.0204003>
- [7] Guy, N. N. (2017). *A Recommender system for rental properties (Thesis)*. Strathmore University

UNDERSTANDING ROMANIAN TEXTS BY USING GAMIFICATION METHODS

*Ștefania-Eliza BERGHIA¹, Bogdan PAHOMI¹,
Daniel VOLOVICI¹,*

*¹ „Lucian Blaga” University of Sibiu, Engineering Faculty, Computer Science
and Electrical and Electronics Engineering Department*

Abstract

In recent years, there has been increasing interest in the field of natural language processing. Determining which syntactic function is right for a specific word is an important task in this field, being useful for a variety of applications like understanding texts, automatic translation and question-answering applications and even in e-learning systems. In the Romanian language, this is an even harder task because of the complexity of the grammar. The present paper falls within the field of "Natural Language Processing", but it also blends with other concepts such as "Gamification", "Social Choice Theory" and "Wisdom of the Crowd". There are two main purposes for developing the application in this paper:

- a) For students to have at their disposal some support through which they can deepen their knowledge about the syntactic functions of the parts of speech, a knowledge that they have accumulated during the teaching hours at school
- b) For collecting data about how the students make their choices, how do they know which grammar role is correct for a specific word, these data being primordial for replicating the learning process

Keywords: Natural Language Processing, Gamification

1 Introduction

Natural Language Processing (NLP) offers one way to make human-computer interaction more interesting and accessible.[1]

The present paper describes a different way of tagging the grammatical functions, by using gamification methods. The main idea is for the students to use the developed application to increase their knowledge about the grammatical roles of the words. They will be able to choose from a list of various syntactic functions, the one that they think is truly eligible for a word or group of words they choose from a text. Their answers are stored and compared with the ones of an expert in the domain, in our case a Romanian language teacher. This comparison will result in a score. On the basis of the score, it will be calculated how much of the answers are correct and a trust will be given. All of this is done in a fun way by using gamification elements to increase user motivation. The goal is to use this data, about the choices of a

user, in implementing a system for automatically tagging the grammatical functions in the Romanian language texts.

2 The Romanian Grammar

As far as the Romanian language is concerned, efforts in the field of natural language processing are made in a few academic centres in Romania and the Republic of Moldova. Among the NLP applications developed, we mention the syllabus application for words, automatic flexing applications (FAVR modelling in the Mac-ELU environment, AnMorph system).

2.1 Parts of Speech in the Romanian Language

The grammar of the Romanian language is considered to be a very complex one. When talking about morphology, we actually refer to the classification of words into parts of speech. In Romanian, from a morphological point of view, the speech parts can be classified into ten types: noun, adjective, pronoun, article, numeral, verb, adverb, preposition, conjunction, interjection. There are also some parts of speech that can be divided into subtypes:

- pronouns: personal, polite, reflexive (which is separate)
- verbs: auxiliary and two other main types
- the article: determined, indefinite, possessive and demonstrative

2.2 Syntactic Functions within Romanian Grammar

According to GALR 2008, "syntactic functions are classes of terms linked by the same type of relationship and the same substitution class "(GALR 2008 II, p. 9) or "classes of substitutable terms in the same context, in other words, classes of functionally equivalent terms in the same position "(ibidem, p. 10).

The main grammatical roles are the following:

1. The subject
2. The predicate
3. The predicative
4. The attribute
5. The complement

Some of these can be of several kinds. For example, for the complement we have:

- object
- adverbial

Model: *Deodată un vânt mare a izbit ferestrele.* (E. Camilar)

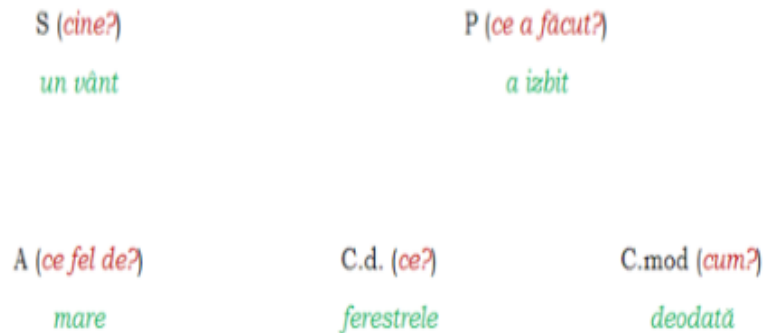


Figure 1: Syntax analysis of a sentence

3 Concept Presentation

In developing this paper we used concepts such as "Gamification", "Social Choice Theory" and "Wisdom of the Crowd".

3.1 1Gamification

In today`s society, students can be considered as digital natives. This means that because they grew up surrounded by technology, they have a different learning process. To keep up with this, the teachers need to adapt this process to the new needs, preferences and requirements of the students in order to maintain their motivation and desire to learn. Such adaptation is exactly the gamification which involves the use of game-specific elements and thoughts, such as giving points based on progress or giving special badges in the context of education.

Game-based approaches lead to a higher level of commitment and motivation of users towards the activities and processes they are involved in. Most of the consumers played or continue to play different games so game mechanics are familiar to them. These approaches based on games are not only true for companies and their employees but in education too.[1] That is precisely why we thought that implementing an application that uses some gamification methods will help the students to learn, test or deepen their knowledge. In this way, their engagement and motivation are enhanced.

3.2 Social Choice Theory

In the Romanian Language analyzing a word or group of words is not an easy task as everything depends very much on the context. In this application, the syntactic analysis of the sentence parts will be done using the Social Choice Theory concept. Basically, this theory explains how to make a decision in the case of voting. We know that elections are based on the majority rule, which classifies a candidate x above a candidate y if and only a majority of individuals do the same.[2]

Let`s see how this applies to our context. For example, for a specific word from a text, the "winning" syntactic function will be considered the one which was chosen most often.

3.3 Wisdom of the Crowd

This phenomenon has gained increasing attention lately. The idea of this is simple: two heads think better than one, and the more they are, the better. It uses the principle of group think, and the concept that the masses are better problem solvers, forecasters, and decision makers than any one individual. An example would be the well-known contest "Who wants to Be a Millionaire". It has been noticed that when one of the contestants was challenged by a particular question, aggregating all the opinions of the people who were left to vote resulted in the correct answer, even if some individuals did not respond properly.[3]

4 Application Implementation

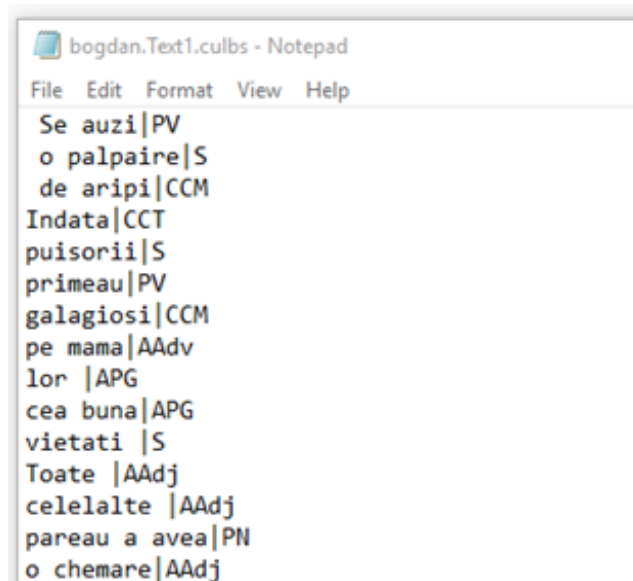
The application from this paper is actually a website. It has a registration/login page, the main page, and a user profile page. In order for the user to access the main page, he will first need to create an account and then log in using the data for which he opted when registering his account. Once logged in, the user will be presented with the main page where he will have certain texts that he will be able to analyze. Also on the main page, the user will be able to view his own profile.

Firstly, for being able to analyze a text the user will need to choose one from a list he will be presented. After that, he will be able to split the text into sentences and take one or more words at a time and choose the syntactic function that seems the more eligible to them, from a list of given grammar functions. When a user finished up analyzing he will submit his answers that will be stored in a database. The answers will be automatically compared with the ones of an expert in domain and a score will be computed. Based on the scores a user gets he will be awarded some merit badges to increase the motivation to learn. He will be able to see his progress and the badges he received in the View Profile section. Also, the scores a user gets are important when computing his "trust rate". The rate will be used to decide if the user is

trustworthy or not. If he gets a high rate it means that his answers were mostly good and we can take them into consideration in a higher proportion. This information among other ones, like the time a user thinks between assigning grammar roles, are essential in replicating the learning process of the students.

5 Results

When submitting the analysis of a text, a file is also created. In this file, we can see the words or group of words and their assigned syntactic functions. This is basically a file with the tags.



```
bogdan.Text1.culbs - Notepad
File Edit Format View Help
Se auzi|PV
o palpaire|S
de aripi|CCM
Indata|CCT
puisorii|S
primeau|PV
galagiosi|CCM
pe mama|AAdv
lor |APG
cea buna|APG
vietati |S
Toate |AAdj
celelalte |AAdj
pareau a avea|PN
o chemare|AAdj
```

Figure 2: Example of a tags file

These files have major importance since they can be used further on, along with the information about how the students assigned the syntactical function to set up a mechanism for understanding Romanian texts.

6 Conclusions and Future Work

This paper presents an application that may be used to analyze texts and also collect data on how the users made their choices. The analysis is made using some gamification methods, like awarding badges and accumulating scores. After the students analyze some texts the information gathered from them and the tags files resulted are to be used in the next step of the project.

As no application is perfect this one is no exception and it can use a series of further improvements and developments such as:

- adding more gamification elements (for example, a user may not be allowed to access some more complex texts until he managed to correctly analyze a series of x texts)

- giving the possibility of analyzing multiple texts
- using collected data to implement a syntactic recognition system for Romanian texts

7 References

- [1] Gabriela Kiryakova, Nadezhda Angelova, Lina Yordanova, "GAMIFICATION IN EDUCATION".
- [2] K. Arrow, „The Principle of Rationality in Collective Decisions” (1952), apud A. Sen (2002: p. 328)
- [3] Aidan Lyon, Eric Pacui, „The Wisdom of Crowds: Methods of Human Judgement Aggregation”, University of Maryland, College Park.

DBSCAN ALGORITHM FOR DOCUMENT CLUSTERING

Radu G. CREȚULESCU¹, Daniel I. MORARIU¹, Macarie BREAZU¹, Daniel VOLOVICI¹,

¹ „Lucian Blaga” University of Sibiu, Engineering Faculty, Computer Science and Electrical and Electronics Engineering Department

Abstract

Document clustering is a problem of automatically grouping similar document into categories based on some similarity metrics. Almost all available data, usually on the web, are unclassified so we need powerful clustering algorithms that work with these types of data. All common search engines return a list of pages relevant to the user query. This list needs to be generated fast and as correct as possible. For this type of problems, because the web pages are unclassified, we need powerful clustering algorithms. In this paper we present a clustering algorithm called DBSCAN – Density-Based Spatial Clustering of Applications with Noise – and its limitations on documents (or web pages) clustering. Documents are represented using the “bag-of-words” representation (word occurrence frequency). For this type o representation usually a lot of algorithms fail. In this paper we use Information Gain as feature selection method and evaluate the DBSCAN algorithm by its capacity to integrate in the clusters all the samples from the dataset.

Keywords: Document Classification, Information Gain, Naive Bayes, Weka framework

1 Introduction

As storage capacity increases, the amount of information saved increases too and become more and more difficult to retrieve and use the saved information. We need more powerful methods that become capable to process this huge quantity of information and offer us an easy and fast access to this information. The text document clustering problem is a special case of an unsupervised learning process in the data mining problem. In order to solve a text document clustering problem some steps are required. The common steps are [6]: feature extraction, feature selection, grouping, evaluation and visualization. The WEKA [9] is a framework that helps us with all these steps. WEKA was initially developed as a library of java classes that help us to implement data mining applications. In the last years, in order to avoid java programming skills, the components from WEKA are also available into a visual form inside the “WEKA Knowledge Flow Environment”..

2 Experimental framework

2.1 Dataset

In order to make some experiments to validate the algorithm functionality we use a Reuters dataset [8], that is a collection of news published by Reuters agency into an XML format, that is close to a text file format. For analyzing and evaluating the learning algorithm we use the Weka learning application that has a lot of learning algorithms already implemented and prepared to be used. In the first part we need to convert the Reuters XML files into a format accepted by Weka, and we need to make lot of steps (classical text mining preprocessing steps [4],[5]) as word extraction, eliminating the common words, keep only the stem of the word and feature selection. We have represented the documents in a vector space model as frequency of word occurrences in document. All the preprocessing steps for transforming the dataset from XML into a Weka format (called arff format) were done into a proper implemented java application. In the resulting file each document is represented on a line as a vector of 1000 different attributes (words). Because Weka has some problems using large numbers of vectors and vectors having a great number of elements, we decide to use only 542 different samples (documents) represented by most relevant 1000 features. The arff format that must be applied to WEKA contains a list with all attributes used into the dataset (defined with name and type as in the next example) and after the "@data" directive it contains a list of each sample (one on a line) with values for each attribute. We prefer the last attribute to be the class (yes/no - if the document is assigned into a class or not). Each sample is classified into a single class (most relevant from the Reuters proposed topic) and we decide to learn only one class. So, for all samples from the dataset that are in that specific class we write 'yes' and for the rest of the samples we write 'no' for the class attribute.

The format for the arff file is as follows:

```
@relation Reuters
@attribute 'A0' numeric
@attribute 'A1' numeric
...
@attribute 'A998' numeric
@attribute 'A999' numeric
@attribute 'class' {'yes', 'no'}
@data
2,2,1,1,1,...,0,0,0,0,0,no
0,0,1,0,1,...,0,2,0,1,0,yes
...
```

In the clustering algorithm we don't need to use the class value that are present in the file in the learning step. We use those values in the feature selection step in order to apply evaluation method for obtain most relevant attributes.

2.2 Weka

For training and evaluating the presented dataset we use the Weka KnowledgeFlow Environment [9],[3] and the experiment uses 6 Weka modules as shown in Figure 1.

The 'ArffLoader' module permits us to load the prepared arff file with the Reuters represented documents. Even if we select the most relevant 1000 features in the document representation step, in this experiment we use a different number of features between 100 to 1000 for evaluating the accuracy of the DBScan algorithm. For modifying the number of features we use a 'Attribute Selection' module that has implemented 'Information Gain' and a method for selecting relevant features.

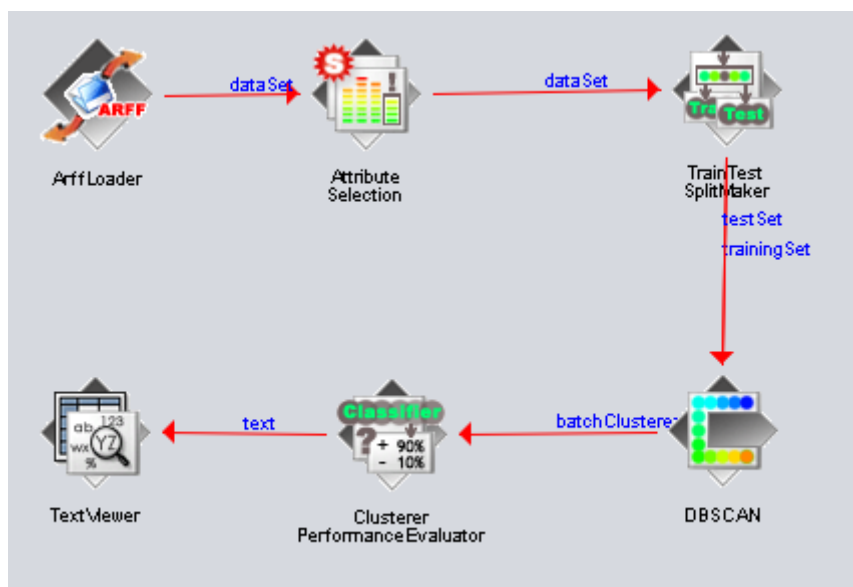


Figure 1. DBScan Experiment in KnowledgeFlow Weka environment

After selection we use a 'TrainTestSplitMarker' that splits the dataset in two parts, one part for training that has 66% of the dataset and the rest for the test.

The 'DBScan' module contains the algorithm that we want to evaluate in this article for the document clustering context. We have some parameters to be specified before running the algorithm: the minimum number of points, the method for distance computing and the value for epsilon (as in figure 2).

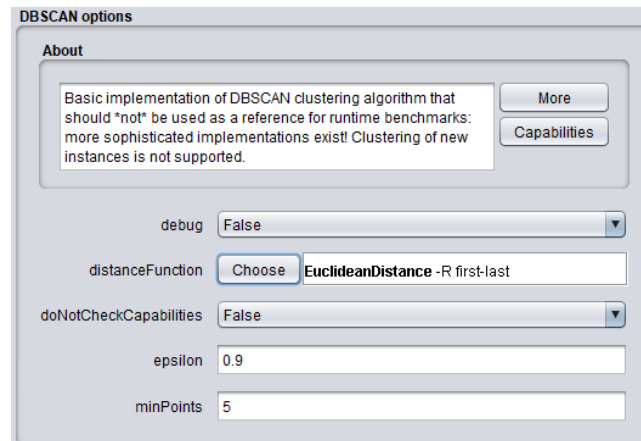


Figure 2. The DBScan configuration

The two last modules are for the cluster evaluation step. We evaluate the algorithm regarding the number of elements present in each class and the needed training time.

3 The clustering algorithm

3.1 DBSCAN (Density-Based Spatial Clustering) Algorithm

DBSCAN [4],[5] is a clustering algorithm based on finding *core objects* and creates a number of groups that are equal with the number of core objects found. A core object is an object that has dense neighborhoods. If the algorithm finds some objects with no dense neighborhoods these are considered noise. It connects core objects and their neighborhoods in order to form dense regions as clusters.

For the configuration of the application we need to be specify the parameter $\epsilon > 0$ for the radius of a neighborhood that is considered for every object. The ϵ -neighborhood of an object o is the space within a radius centered in o . The density of a neighborhood can be measured by the number of objects in the neighborhood. Another user-specified parameter is *MinPts*. This parameter specifies the minimum density threshold for a dense region to be considered as core object. An object is considered to be a **core object** if the ϵ -neighborhood of the object contains at least *MinPts* objects.

For a core object q and an object p , we say that ***p is directly density-reachable from q*** if p is within the ϵ -neighborhood of q . In figure 3 are represented al symbols.

Density reachability respects following rules:

- p is density-reachable from q (with respect to ϵ and *MinPts* in dataset) if there is a chain of objects p_1, \dots, p_n , such that $p_1 = q$, $p_n = p$, and p_{i+1} is directly density-reachable from p_i with respect to ϵ and *MinPts*;

- if o_1 and o_2 are core objects and o_1 is density-reachable from o_2 , then o_2 is density-reachable from o_1 ;

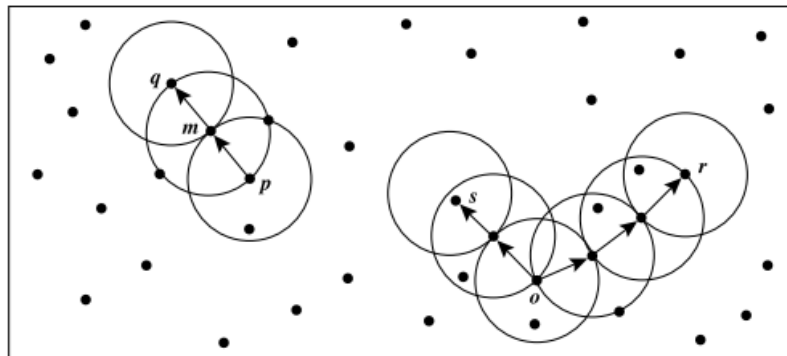


Figure 3 Density-reachability and density-connectivity (3)

- if o_2 is a core object but o_1 is not, then o_1 may be density-reachable from o_2 , but only in this direction.

Density connectivity respects following rules:

- p_1, p_2 are density-connected with respect to ε and *MinPts* if there is an object q such that both p_1 and p_2 are density-reachable from q with respect to ε and *MinPts*;
- Unlike density-reachability, density-connectedness is an equivalence relation. If o_1 and o_2 are density-connected, and o_2 and o_3 are density-connected, then also o_1 and o_3 are density-connected.

The density-connectedness is used to find connected dense regions as clusters. Each closed set is considered to be a density-based cluster. A subset of elements from dataset is considered a cluster if it satisfies:

- for any two objects o_1, o_2 are considered to be clusters if o_1 and o_2 are density-connected;
- there does not exist an object o that is a cluster and another object o' that is not a cluster but o and o' are density-connected.

The advantage of the algorithm is that it achieves all the time same results but has a huge disadvantage because it makes nothing to represent the training dataset into a simplified manner. It uses all the time the entire dataset, that means the testing part is time consuming.

3.2 Evaluating clustering performance

The evaluation of the clustering algorithm is a problem because the groups are created automatically by the algorithm, from algorithm point of view, and don't corresponds with the class saved in the dataset. For cluster evaluation Weka [9] offers three different methods depending on the selected clustering method:

- *Use training set* (this is the default mode), when after generating the cluster Weka will classify the training instances into clusters according to the cluster representation and computes the percentage of instances assigned to each cluster;
- *In supplied test set or Percentage split*: it is used a separate test dataset for evaluating the clustering. This works only if the cluster representation is probabilistic;
- *Classes to cluster* where Weka first ignores the class attribute and generates the clusters. Then in the test phase it assigns classes to the clusters based on the majority value of the class attribute within each cluster. After that the classification error can be computed, based on this assignment.

Because we have only one class in the dataset, we use the default mode used by Weka to evaluate the algorithm. In this case we evaluate looking at the number of elements that remain outside of the clusters (numbers of samples which the algorithm didn't put into a cluster (these are considered noise in the dataset)).

4 Experimental results

We intend to evaluate the algorithm regarding the number of attributes used in the learning part and regarding the influence of the algorithm parameters. Because we use a clustering algorithm, we don't use the class attribute present in the dataset. That class is used only in the feature selection method, for select the best attributes. This algorithm generates a different number of classes, depending on the parameters. We evaluate the algorithm based on the number of samples that are considered to be noise and are not grouped.

4.1 Influence of input parameters

For all the features from the initial file (1000 features) we evaluate the number of elements that are left out in the learning step and in the testing step. We have 542 samples that were randomly split by Weka in a training set with 358 samples and in the testing set with 184 samples. The algorithm generates between 3 and 7 clusters.

For the 1000 attributes we vary the ϵ between 0.9 and 0.1 and keep the *MinPts* constant. The number of documents in the resulted clusters for the training set are presented in Table 1 (358 samples and 1000 attributes)

Table 1. Learning rate for training dataset

ϵ	MinPts	Nr. clusters	C1	C2	C3	C4	C5	C6	C7	Noise	Learning Rate
0,9	6	3	102	61	19					176	50,84%
0,8	6	3	100	56	19					183	48,88%
0,7	6	3	94	49	19					196	45,25%

0,6	6	3	94	45	19					200	44,13%
0,5	6	6	41	7	27	7	11	8		257	28,21%
0,4	6	6	41	7	27	7	11	8		257	28,21%
0,3	6	6	40	7	25	7	11	8		260	27,37%
0,2	6	7	25	6	16	6	10	6	8	281	21,51%
0,1	6	6	10	6	15	6	13	7		301	15,92%

The column named 'Learning Rate' presents the percent of elements that were grouped by the algorithm into clusters. When this value is small it means a higher number of elements that were not grouped and were considered noise (these are presented also as number in column 'Noise'). From this point of view we see that a small value for ϵ leads to an increased number of noise elements in the results.

For the testing set we present the results in the Table 2 (184 samples and 1000 attributes). In the testing part the algorithm will cluster the samples in learned groups. For the training and testing part we use the Euclidean distance as measure for computing the distance.

Table 2. Learning rate for testing dataset

ϵ	MinPts	Nr. clusters	C1	C2	C3	C4	C5	C6	C7	Noise	Learning Rate
0,9	6	3	48	27	10					99	46,20%
0,8	6	3	47	26	10					101	45,11%
0,7	6	3	57	24	10					105	46,43%
0,6	6	3	45	22	10					107	41,85%
0,5	6	6	18	4	16	5	4	4		133	27,72%
0,4	6	6	18	4	16	5	4	4		133	27,72%
0,3	6	6	18	4	16	5	4	4		133	27,72%
0,2	6	7	13	3	11	4	4	2	4	143	22,28%
0,1	6	6	5	3	10	4	6	4		152	17,39%

When we decrease the ϵ value the number of elements that are considered noise will increase in the same way for the training and for the testing dataset. It's interesting that when we create more clusters the number of documents that remain outside increases too.

In table 3 we present for the training dataset the influence of *MinPts* (number of points that are needed to be taken into consideration to create a core object and a cluster). Because we have obtained best values for $\epsilon=0.9$ we will present the experiments only for this value.

Table 3. Influence of *MinPts* for training dataset

ϵ	MinPts	Nr. clusters	C1	C2	C3	C4	C5	C6	C7	...	Noise	Learning Rate
0,9	2	11	2	102	61	19	2	2	2	...	158	54,60%
0,9	3	3	104	62	19						173	51,68%
0,9	4	4	102	61	19	4					172	51,96%
0,9	5	3	102	61	19						176	50,84%
0,9	6	3	102	61	19						176	50,84%

0,9	7	3	102	59	19						178	50,28%
0,9	8	3	103	61	19						175	51,12%
0,9	10	3	101	59	19						179	50,00%
0,9	15	3	101	59	19						179	50,00%
0,9	20	2	101	59							198	44,69%

When we have a small number of *MinPts* the algorithm creates a large number of clusters. In table 4 we present only the values for the first 7 clusters. When the number of *MinPts* increases the number of clusters decrease but the number of elements that are considered noise increases too.

Table 4. Influence of MinPts for test dataset

ϵ	MinPts	Nr. clusters	C1	C2	C3	C4	C5	C6	C7	...	Noise	Learning Rate
0,9	2	11	1	48	27	10	1	2	2	...	88	50,84%
0,9	3	3	49	27	10						98	46,74%
0,9	4	4	48	27	10	1					98	46,74%
0,9	5	3	48	27	10						99	46,20%
0,9	6	3	48	27	10						99	46,20%
0,9	7	3	48	26	10						100	45,65%
0,9	8	3	48	27	10						99	46,20%
0,9	10	3	47	26	10						101	45,11%
0,9	15	3	47	26	10						101	45,11%
0,9	20	2	47	26							111	39,67%

From the 'Noise' point of view we can observe that, when the number of *MinPts* increases, for the same value for ϵ , the number of considered noise samples increases also. This means that in the file the samples are not uniformly distributed, and this becomes a problem to group such type of samples. Also, because we can have a relatively a small number of samples, we can't increase the *MinPts* value. The conclusion is that for text documents we need to use a large value for ϵ and a small number for *MinPts*.

5 Conclusions

The purpose of this paper was to analyze the performance of the DBScan algorithm in the context of clustering text documents. For this we have used the Weka implementation for the DBScan algorithm and algorithm evaluation, and we have used our own implementation for processing the Reuters files and create the dataset that respects the input Weka format.

We have evaluated the clustering algorithm from the point of view of the number of samples (Noise) that are left outside in the training and testing dataset. We have considered that if this value increases, the quality of the learning algorithm decreases. As it can be observed, we obtain the best values, in case of text document clustering, for a small number of *MinPts* (points that need to be taken in consideration for creating a core object) and

for a high value for Epsilon (radius of a neighborhood that is considered for each object).

6 References

- [1] S. Chakrabarti, *Mining the Web- Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Press, 2003;
- [2] Cretulescu, R., Morariu, D. *Text Mining. Tehnici de clasificare si clustering al documentelor*, Published at Editura Albastra, Cluj Napoca, 2012, ISBN 978-973-650-289-7
- [3] Radu Cretulescu, Daniel Morariu, Macarie Breazu - *Using WEKA framework in document classification*, Int. Journal of Advanced Statistics and IT&C for Economics and Life Sciences, Vol 6, No 2, ISSN 2067-354X, 2016
- [4] Han, J., Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001;
- [5] Mitchell T. *Machine Learning*, McGraw Hill Publishers, 1997.
- [6] Mitkov R., *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2005;
- [7] Daniel Morariu, Radu Cretulescu, Macarie Breazu - *The WEKA Multilayer Perceptron Classifier*, Int. Journal of Advanced Statistics and IT&C for Economics and Life Sciences, Vol 7, No 1, ISSN 2067-354X, 2017
- [8] Reuters Corpus: <http://about.reuters.com/researchandstandards/corpus/>. Released in November 2000
- [9] WEKA package - <http://www.cs.waikato.ac.nz/ml/weka/index.html> (accessed March 2015)

International Journal of Advanced Statistics and IT&C
for
Economics and Life Sciences

VOLUME 9 Number 1 CONTENTS 2019

LIBRARY AND INFORMATION SCIENCE VIS-À-VIS WEB SCIENCE IN THE LIGHT OF THE OECD FIELDS OF SCIENCE AND TECHNOLOGY CLASSIFICATION M. Grabowska	3
STUDY ON THE MAPPING OF RESEARCH DATA IN THE REPUBLIC OF MOLDOVA IN THE CONTEXT OF OPEN SCIENCE N. Turcan, A. Rusu, R. Cujba	13
DIGITAL RISKS. CASE STUDY ON DIGITIZATION PROJECTS OF THE LBUS LIBRARY R.,M. Volovici, E. Mărginean, Io.Vișa	25
THE EUROPEANA COLLECTIONS – TRANSYLVANIAN PRINTS FROM THE COLLECTIONS OF THE NATIONAL MUSEUM OF THE UNION FROM ALBA IULIA A. Roman Negoii	33
SUGGEST RECOMMENDATION FOR LIBRARY USERS USING GRAPHS Gh., C. Crișan	43
UNDERSTANDING ROMANIAN TEXTS BY USING GAMIFICATION METHODS Șt. Berghia, B. Pahomi, D. Volovici	55
DBSCAN ALGORITHM FOR DOCUMENT CLUSTERING R. Crețulescu, D. Morariu, M. Breazu, D. Volovici	61

ISSN 2067 – 354X