e-mentor

DWUMIESIĘCZNIK SZKOŁY GŁÓWNEJ HANDLOWEJ W WARSZAWIE

WSPÓŁWYDAWCA: FUNDACJA PROMOCJI I AKREDYTACJ KIERUNKÓW EKONOMICZNYCH

2019, No 3(80)

Gmiterek, G. (2019). Challenges Related to Identifying Sources and Document Collection for Big Data Analyses. *e-mentor*, *3*(80), 4–9. DOI: 10.15219/em80.1417

Challenges Related to Identifying Sources and Document Collection for Big Data Analyses

Grzegorz Gmiterek[*]

For some time now we have been witnessing a gradual but dynamic increase in the volume of information, the existence of which has had an increasing effect on our lives. The amount of digital data generated daily is estimated at 3 trillion bytes, and continues to grow by the day (Global AI Bootcamp, 2018). This is a result of data being generated by various digital platforms, as well as sensors, the internet and the billions of cell phones used every day (Henke et al., 2016). The mass storage capacity of these devices is increasing, and the cost of storing data is declining rapidly. The information overflow which occurred during the first decade of the 21st century was accompanied by phenomena such as the rapid influx of information and knowledge, but also the intensification of information chaos (Babik, 2010, p. 22). The creation of mass information resources is a global phenomenon affecting various areas of life, science, culture, the economy, etc. It is thus no wonder that Płoszajski (2013, p. 6) claims that we are witnessing the advent of a revolution which brings with it unlimited data processing as a result of the exponential growth in computing power, as well as broader and more affordable access to the data.

For the purpose of this article, the following definition of big data found in the relevant literature is used: *the term is used to refer to data sets which are large in volume and diverse, accumulate in real time, are changing, complex and require the use of innovative technologies, tools and IT methods for the purpose of extracting new and useful knowledge* (Tabakow, Korczak and Franczyk, 2014, p. 141).

Purpose of the article and methods

The purpose of this article is to analyze selected issues and challenges related to identifying and collecting large volumes of digital information (big data) from the perspective of information science and within the context of applying the tools and resources offered by libraries (primarily research libraries). The research process was qualitative in nature, with a critical analysis of the relevant literature as well as of documents found in technical and specialist press and on websites. An important source of knowledge on publications concerned with identifying and collecting large volumes of data was the Library, Information Science & Technology Abstracts database, which contains information on sources from the areas of book science, information science, bibliometry etc. A query was also conducted in the abstract and full-text databases available via the Chamo discovery system offered by the Warsaw University Library, which integrates the resources from several dozen scientific and thematic sources.

The purpose of the article was also to characterize certain functional tools (in particular discovery systems) and to identify major issues related to sources and searches in the processing, collecting and organizing of large volumes of data, as well as accessibility and classification from the perspective of document typology. An important source of knowledge on the topic was the experience acquired by the author over the course of the research project *Exploring sources of data on the topic of R+D+I activity*, financed via a grant awarded by the National Center for Research and Development and conducted at the Department of Media Information Technologies of the Faculty of Journalism, Information and Book Studies, University of Warsaw. The project was carried out between July 2017 and December 2018 under the supervision of Professor Włodzimierz Gogołek.

Due to article size restrictions, only selected issues have been analyzed in this paper. Not all major issues have been taken into account that a discerning reader could consider meriting analysis (e.g. the typology of sources and search tool classification criteria, information search strategies, the applications and functionalities of tools facilitating automated digital resource collection, using software which adds new functionalities to search engines and browsers). Hence a comprehensive analysis of the issues related to big data source identification, although it would be of much methodological utility, is beyond the scope of this article. However, the author hopes that this topic will be analyzed more extensively in the future.

[*] University of Warsaw, Poland

Information science and libraries in the world of big data processing

Big data processing has been at the center of much debate across research publications from various fields. The Web of Science (an abstract and bibliometric database developed by Thomson Reuters) database shows an increase in the number of publications submitted on this topic starting from 2012. These are publications from the fields of computer science, engineering, telecommunications, business, the social sciences, mathematics, medicine, education, environmental protection and information science. In the case of the latter, the major issues appear to be identifying sources, as well as searching, collecting, storing and organizing the large volumes of data, which requires the use of unconventional methods and tools in order to access them (including data located in what is referred to as the deep web). Of utility in this respect is an understanding of effective search methods and strategies, query languages and effective use of the bibliographic, abstract and full-text databases offered by institutions such as libraries. This applies particularly in situations where the large volume of data to be analyzed consists primarily of scientific sources.

Issues related to big data analysis have been discussed in book and information science literature, including articles on the functioning of libraries, methods of organizing resources and information/library processes. Information science is a discipline *concerned with the theoretical and practical aspects of information-related activities, their tools and methods, including, in particular, information systems, informative sources, information sources, methods and tools used in information processes and the information-related behaviors of users of information systems, sources and services* (Sosińska-Kalata and Pindlowa, 2017, p. 725).

It should be noted that Nicholson coined the term *bibliomining* as early as in 2003, referring to the use of data exploration methods by libraries (Nicholson, 2003, p. 146). However, it was not until recently that a marked increase in publications on issues such as big data processing has been observed within the context of information and database management. Even a cursory analysis of the most relevant search results in the Library, Information Science & Technology Abstracts database demonstrates that there exist numerous publications on big data related to such subjects as social media, IT, AI, digital publications, data exploration and the related software, ontology, metadata and digital libraries.[1] A large proportion of these are publications presenting the idea of big data, but also indicating the challenges and possibilities related to the phenomenon. The issue is also being discussed with increasing frequency on expert portals and blogs (Wójcik, 2016, p. 63), while librarians and information scientists discuss the topic in social media.[2] That there is interest in the topic is evident by the increasing popularity of conferences organized in Poland on issues that relate to varying degrees with the processing of large volumes of data.[3]

Information scientists also describe the issue of developing a methodology to make it possible to effectively identify trends (Świgoń, 2018; Ahmed and Ameen, 2017), as well as the sources of documents and publications for big data processing (Ball, 2013; Cuzzocrea, Simitsis and Song, 2017; Hoy, 2014; Reinhalter and Wittmann, 2014; Tuppen, Rose and Drosopoulou, 2016). In this case, Świgoń (2018, p. 128) defines trends as *primarily technological changes, but also other changes, e.g. social, cultural, environmental, economic, and political, which influence the human infosphere*. The infosphere is also referred to by information scientists as information space, information environment or information networks (Ibid.).

It appears that the issue of identifying trends is to a large extent related to the application of tools, abilities and methods used every day by library employees (especially employees of research and university libraries). The issue is also related to library tools which integrate the ability to search diverse digital resources, which may be part of licensed abstract and full-text bibliographic databases, abstract and bibliometric databases, as well as digital libraries and repositories. The tools offered by libraries constitute one of the fundamental methods of verifying content, which comprises frequently updated, selected literature on a given research problem. It is important to remember that this literature is not only limited to scholarly analyses (expert knowledge on a subject). Libraries, especially research and specialized libraries, offer a range of tools which enable access to various types of data, including what is referred to as gray literature.

[1] Based on the results of a search query conducted in the Library, Information Science & Technology Abstracts database on 12.21.2018. The purpose of the search was to find articles related to the term "big data". Information was found on 1262 documents created between 2010 and 2018.

[2] An example of this is Facebook, where a number of posts can be found by librarians and information scientists from around the globe.

[3] Recent examples of such conferences include: *Big Data w humanistyce i naukach społecznych* (*Big data in the humanities and social sciences,* 21–22 November 2018, Institute of Library and Information Science of the University of Wrocław); XII edition of the *Automatyzacja Bibliotek* (*Library Automation*) series: *Biblioteki w cyberprzestrzeni: Inspiracje Światowego Kongresu IFLA Wrocław 2017* (*Libraries in Cyberspace: Inspirations of the 2017 IFLA World Congress in Wrocław,* 13–14 November 2018, Warsaw); *Małopolskie Forum Bibliotek. Biblioteka 2.028. Między nadmiarem a niedostatkiem* (*Lesser Poland Library Forum. Library 2.028. Between excess and deficiency,* 24–26 October, Krakow); *VIII Krajowa Konferencja Naukowa INFOBAZY 2017 – Bazy Danych dla Nauki – dane, wiedza, e-usługi* (*2017 INFOBASES VIII National Scientific Conference – Databases for Science – data, knowledge, e-services,* 11–13 September 2017, Gdansk).

Nevertheless, the knowledge on information science possessed by librarians and information specialists may also be of utility, including the knowledge of internet search engine trends, methods of determining search and data collection strategies and sources of various types and applications. The claim expressed in an Onet.pl interview by Wiesław Cetera, one of the researchers responsible for the project *Exploring sources of data on B+R+I activity*, according to which "data processing is at the center of attention during design work and is a pivotal part of the system used in the project, but it all begins with people – information scientists who can steer exploration in the right direction," thus appears valid. "They determine where we can expect to find useful data. It is necessary to reach them" (Żemła, 2018). Naturally, depending on the research problems and areas in which a user would like to explore sources, a different search and identification strategy will be implemented and other tools will be used. This may constitute a heuristic method of source identification and document searching.

Discovery systems and identifying document sources for big data analyses

From the perspective of identifying research articles or documents, collecting them for further big data analyses, as well as using accurate and up-to-date data for determining long-term development trends, it appears important to use the functionalities offered by tools referred to as discovery systems. These enable quick discovery and convenient retrieval of sources available in a given library (or library group). Discovery systems enable integrated searches in multiple sources at one time, e.g. full-text databases, bibliographic databases, bibliographic and abstract databases, digital libraries and archives, library catalogs, as well as blogs, websites and themed resource guides – LibGuides.[4] Example institutions which make extensive use of the above include the Villanova University Library in the USA[5] and the Oxford University library.[6]

As mentioned earlier, discovery systems are currently also used to develop tools which integrate information on sources from several institutions, both traditional and digital. The Lublin Virtual Library is one such case, as it enables users to search library catalogs of databases, commercial full-text bibliographic and abstract databases, as well as open access resources. The project was developed based on the Primo system created by ExLibris.[7] It is worth noting that discovery systems are analyzed in the relevant literature both within the context of discovery interfaces and discovery services. The solution in question constitutes software which is installed in addition to the document information access tools offered by a given library and utilizes cloud computing (Skórka, 2017, p. 136). Modern discovery systems include such open source solutions as VuFind (used by the aforementioned Villanova University library) and commercial solutions like Primo ExLibris and Chamo Discovery.[8]

A distinctive feature of the above systems is that they offer users advanced search options and faceted management of relevant search results. Facets in this case refer to the attributes of the searched documents. Using facets can help narrow down search results within a single category, organized according to a criterion available in the faceted navigation function, such as author, collection, language, type of document, date etc. As emphasized by Laura Morse from the Open Discovery Initiative and Harvard University, *discovery systems are of extreme significance to the research ecosystem* (NISO, 2014). They enable users to find the most up-to-date sources of knowledge created by researchers from around the world. What is more, such tools constitute a fundamental method of verifying content, which comprises frequently updated, selected literature on a given research problem and the accompanying metadata.

It is also worth noting that individual bibliographic, bibliographic-abstract and full-text databases also offer the aforementioned faceted navigation function. In addition, some of these databases (e.g. Science Direct) offer the option to bulk download entire texts or export search results as publication metadata into a calculation sheet or bibliography manager tool (e.g. IEEE Xplore Digital Library enables exporting search results containing the detailed metadata of publications and their abstracts as CSV files).

Typology of identified documents

Within the context of identifying large volumes of data it is worth giving at least a cursory review of the issue of the typology or organization of the documents to be processed. A suitable typology may render it easier to organize the collected content and the subsequent data analysis and result interpretation. Articles to be processed come from various sources,

[4] More information on LibGuides can be found in Derfert-Wolf, L. (2011). Specjalista informacji 2.0? Bibliotekarz dziedzinowy 2.0? Nowa forma przewodników po zasobach – LibGuides. *Biuletyn EBIB*, *119*(1). Retrieved from http://tiny.cc/4grhcz

[5] Villanova University. (n.d.) *Falvey*. *Memorial Library*. Retrieved 01.18.2019 from https://library.villanova.edu/

[6] Bodleian Libraries. University of Oxford. Search Oxford's Libraries Online. (n.d.) Retrieved 9.08.2019 from https://www.bodleian.ox.ac.uk/

[7] Lubelska Biblioteka Wirtualna. (n.d.) Retrieved 9.08.2019 from http://projektlbw.lublin.eu/

[8] More information on discovery systems, their features and functionalities can be found in the doctoral dissertation of Dominika Paleczna. The dissertation is available in the electronic repository of Warsaw University. Paleczna, D. (2016). *Aspects of Designing and Evaluation of Library Information Systems at the beginning of the 21st century*. Retrieved from https://depotuw.ceon.pl/handle/item/1531

e.g. blogs, microblogs, forums, information portals, RSS channels, data wholesalers (see Busłowska and Wiktorzak, 2014, pp. 2491–2493; Drosio and Stanek, 2017, p. 107; Maślanowski, 2015, p. 168; Rodak, 2017; TQMSOFT, 2018). The data typically processed today come from various meters, sensors, logs, GPS devices, as well in the form of sequences of clicks on websites (Burnet-Wyrwa, 2017, p. 47). These may to a large extent constitute open-access data (Otwarte Dane, n.d.).

Resources which are to be processed may be external or internal, historical or current. The diversity of documents in big data analyses is thus a key issue, particularly in the case of processing unstructured data sets. To be more precise, it is important to note that currently more than 90% of information is recorded in an unstructured form (Kim, Trimi and Chung, 2014, p. 78).

The document typology used by a researcher can be based on content (e.g. industry-related, research, education, official, social), form of presentation, file format, method of access (including the degree of openness and confidentiality, availability in the surface and deep web), relevance, completeness, multimedia use and being up-to-date with regard to the information contained. In the case of identifying resources for big data processing, the classical typology can also be applied, which refers to the traditional division of sources into primary, secondary and derivative. This typology takes into account the origins of sources, the way they were created and the degree to which they have been processed. A distinctive feature of primary sources is that their content remains in the original form envisioned by the author. Secondary sources are created based on primary or derivative sources and reflect their features in terms of content (Głowacka, Jarocki, Kowalska, Kurkowska and Pamuła-Cieślak, 2016, p. 190). Derivative sources contain information on primary and secondary sources. Their distinctive feature is that they enable readers to learn about the content of the other two types of sources (Ibid.). Examples of derivative sources include bibliographic descriptions and bibliographies, documentation descriptions and thematic compilations (Hancko, 1972, pp. 18–26). However, in today's world, these are also abstract databases and publications, RSS channels, information bulletins, guides and other types of sources.

Open access to resources and collecting large volumes of data

One of the purposes of research libraries is to collect high-quality library resources. This implies that such institutions should create spaces where knowledge is collected and made available as part of historical collections, as well as being contained in the latest, up-to-date information sources (Materska, 2016a, p. 65). However, it should be noted that, in the case of digital sources available via libraries, the number of documents possible to be downloaded by a user for the purpose of further processing is frequently limited. In the case of license-based databases offered by libraries, their terms of service frequently clearly specify how many sources can be downloaded (regardless of whether a discovery system, individual database interfaces or other library resources are used). Importantly, such terms of service may be relatively restrictive, depending on the library and the type of sources it offers. An example of this is the website of the Warsaw University Library, which contains a provision which states that it is permissible to download "PDF and other available files using the publisher's platform functions, but only for personal, private, scientific, educational and research use"(Warsaw University Library, n.d.). The "mass and automatic (with the use of software) downloading of files and other data" is prohibited (Ibid.).

Open access to digital resources appears to be of high importance in modern big data analyses; for example, within the context of the content of a large portion of digital libraries and repositories, as well as certain full-text and abstract databases. Naturally, the above only applies if the terms and conditions of using such sources do not impose restrictions on downloading large amounts of data. It is also important to note that the open access movement emphasizes free access to the results of publicly-funded research. This finds confirmation in the recommendation contained in the 2005 OECD Report, for example, according to which governments should recoup research costs by, among other methods, sharing research results as broadly as possible (OECD, 2005). In addition, institutions which finance research demand with increasing frequency that study results be available on an open-access basis. Such solutions are utilized by the European Commission and the European Research Council, for example. The Horizon 2020 framework program and its guidelines exemplify this approach. Open access, which is the free online sharing of research information and study data, as well as the ability to reuse them in the future, is considered by these guidelines an important aspect of the process of sharing project results (European Commission, 2018). As noted correctly by Materska (2016b, p. 51), *at the current stage the policy for open access to scientific publications is promoted as obligatory in relation to materials and data created in the process of publicly-funded research*.

In light of the above and within the context of using scientific and technical sources for the purpose of big data processing, the concept of open public sector data is also important, as is the ability to reuse such data (e.g. for research purposes). It is worth noting that the concept of sharing public data is in line with the guidelines of the European Commission (Directive 2013/37/EU of the European Parliament and of the Council, 2013; Otwarte Dane, 2018). The idea of openness applies in particular to collecting and sharing data whose creation involved public funding. Such data may be used, processed and published, provided that their source is indicated and that further distribution of content based on that data follows the same rules (Pawełoszek, 2014, p. 456).

Summary

We are currently witnessing a gradual increase in the demand for specialists in big data analysis and processing (e.g. big data scientists, big data architects and big data analysts).[9] This applies to structured, partially structured and unstructured data contained within scattered databases, file collections, portals, libraries, repositories and digital archives, on websites and online forums, relayed via streams, in social media and other online sources. In identifying and collecting resources, certain issues and challenges related to information science can be identified, as well as certain legal and tool-related (including IT) issues. Bibliologists, information scientists and librarians certainly possess knowledge and skills related to specialist sources of information. Moreover, librarians possess the tools to effectively facilitate the process of finding documents which are most relevant to a given search query. Their knowledge of how to identify and collect documents for the purpose of big data processing can in many cases prove invaluable.

References

Ahmed, W., & Ameen, K. (2017). Defining big data and measuring its associated trends in the field of information and library management. *Library Hi Tech News*, *34*(9), 21–24. DOI: https://doi.org/10.1108/LHTN-05-2017-0035

Babik, W. (2010). O natłoku informacji i związanym z nim przeciążeniu informacyjnym. In J. Morbitzer (Ed.), *Człowiek-Media-Edukacja* (pp. 21–27). Kraków: Katedra Technologii i Mediów Edukacyjnych Uniwersytetu Pedagogicznego w Krakowie.

Ball, S. (2013). Managing big data: what's relevant? *AALL Spectrum, 18*(2), 25–27.

Biblioteka Uniwersytecka w Warszawie. (2019). *Zasady wykorzystania baz*. Retrieved from https://tinyurl.com/y5eepvqz

Bodleian Libraries. University of Oxford. Search Oxford's Libraries Online. (n.d.). Retrieved 8.09.2019 from https://www.bodleian.ox.ac.uk/

Burnet-Wyrwa, W. (2017). Big Data – wyzwanie dla rachunkowości zarządczej. *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 341*, 45–53.

Busłowska, E., & Wiktorzak, A.A. (2014). Przetwarzanie i analizowanie dużych ilości danych, *Logistyka*, *6*, 2490–2496.

Cisek, S. (2017). *Podstawy teorii i metodologii informatologii*. Retrieved from https://tinyurl.com/y6lc7ezn

Clarivate Analytics. (2019). *Web of Science. Quick reference guide*. Retrieved from https://tinyurl.com/y4u8qf6c

Cuzzocrea, A., Simitsis, A., & Song, I. (2017). Big Data Management: New Frontiers, New Paradigms. *Information Systems, 63*, 63–65. DOI: https://doi.org/10.1016/j.is.2016.07.002

Derfert-Wolf, L. (2011). Specjalista informacji 2.0? Bibliotekarz dziedzinowy 2.0? Nowa forma przewodników po zasobach – LibGuides. *Biuletyn EBIB*, *119*(1). Retrieved from https://tinyurl.com/y3nbbwe7

Drosio, S., & Stanek, S. (2017). Big Data jako źródło informacji rozszerzające funkcjonowanie systemów wspomagania decyzji w zarządzaniu kryzysowym. *Zeszyty Naukowe Politechniki Częstochowskiej. Zarządzanie, 26*, 107–120.

European Parliament. (2013). *Dyrektywa Parlamentu Europejskiego i Rady 2013/37/UE z dnia 23 czerwca 2013 roku*. Retrieved from https://tinyurl.com/y5l7ky78

European Commission. (2018). *Open Access to scientific information*. Retrieved from https://tinyurl.com/y4tptjrw

Global AI Bootcamp. (2019). Retrieved from https://tinyurl.com/y68ltmsj

Głowacka, E., Jarocki, M., Kowalska, M., Kurkowska, E., & Pamuła-Cieślak, N. (2016). Współczesne źródła informacji. In W. Babik (Ed.), *Nauka o informacji* (pp. 189–214). Warszawa: Wydawnictwo SBP.

Hancko, R. (1972). *Zarys wiadomości o dokumentach*. Warszawa: Centralna Biblioteka Wojskowa.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B.I., & Sethupathy, G. (2016). *Report McKinsey Global Institute. The age of analytics: Competing in a data-driven world. McKinsey Global Institute*. Retrieved from https://tinyurl.com/yb7vytkg

Hoy, M.B. (2014). Big Data: An Introduction for Librarians. *Medical Reference Services Quarterly, 33*(3), 320–326. DOI: https://doi.org/10.1080/02763869.2014.925709

Kim, G-H., Trimi, S.,& Chung, J-H. (2014). Big-Data Applications in the Government Sector. *Communications of the ACM*, *57*(3), 78–85. DOI: 10.1145/2500873

Lubelska Biblioteka Wirtualna. (2019). Retrieved from http://projektlbw.lublin.eu/

Materska, K. (2016a). Biblioteka akademicka jako element infrastruktury naukowej w cyfrowym świecie danych, informacji i wiedzy. In H. Brzezińska-Stec, & J. Żochowska (Eds.), *Biblioteki bez użytkowników…? Diagnoza problemu* (pp. 53–68). Białystok: Wydawnictwo Uniwersytetu w Białymstoku.

Materska, K. (2016b). Aktualność koncepcji zarządzania informacją w dobie big data – perspektywa informatologa. In S. Cisek (Ed.), *Inspiracje i innowacje: zarządzanie informacją w perspektywie bibliologii i informatologii* (pp. 47–59). Kraków: Biblioteka Jagiellońska.

McKinsey & Company, (2016). *Raport o zaawansowanej analizie danych i Big Data*. Retrieved from https://tinyurl.com/y2px289n

Maślanowski, J. (2015). Analiza jakości danych pozyskiwanych ze stron internetowych z wykorzystaniem rozwiązań Big Data. *Roczniki Kolegium Analiz Ekonomicznych, 38*, 167–177.

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries, 22*(4), 146–151.

NISO, (2014). *NISO Launches Open Discovery Initiative (ODI) Standing Committee*. Retrieved from https://tinyurl.com/y5tmtled

[9] Recent job market research demonstrates that there has been a gradual increase in the demand for staff possessing high analytical skills related to data source identification, big data exploration and analysis and using the results of such analyses in decision-making processes in various areas of economic activity (the demand for such specialists may grow by as much as 12% per year, and in the U.S. alone, between 2 and 4 million jobs will be created for such professionals in the following decade) (see McKinsey & Company, 2016).

Organisation for Economic Co-operation and Development (OECD). (2005). *Digital Broadband Content: Scientific Publishing*. Retrieved from https://tinyurl.com/y2a9x3pk

Otwarte Dane. (2019). *Korzystaj z danych!* Retrieved from https://dane.gov.pl/

Otwarte Dane. (2018). *Portal Open Data Gotowy do wdrożenia w Twoim mieście*. Retrieved from https://www.otwartedane.com/

Paleczna, D. (2016). *Aspekty projektowania i oceny systemów informacyjno-wyszukiwawczych bibliotek na początku XXI w.* Retrieved from https://depotuw.ceon.pl/handle/item/1531

Pawełoszek, I. (2014). Wybrane problemy wdrożenia koncepcji otwartych danych w e-administracji. *Roczniki Kolegium Analiz Ekonomicznych, 33*, 455–470.

Płoszajski, P. (2013). Big Data: nowe źródło przewag i wzrostu firm. *e-mentor, 3*(50), 5–10.

Reinhalter, L., & Wittmann, R.J. (2014). The Library: Big Data's Boomtown. *Serials Librarian*, *67*(4), 363–372. DOI: https://doi.org/10.1080/0361526X.2014.915605

Rodak, O. (2017). Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: perspektywa konstruktywistyczna. *Studia Socjologiczne, 3*(226), 209–236.

Skórka, S. (2017). Wizualizacja nawigacji w serwisach typu discovery. *Toruńskie Studia Bibliologiczne, 2*(19), 135–161.

Sosińska-Kalata, & B., Pindlowa, W. (2017). Informatologia. In A. Żbikowska-Migoń, & M. Skalska-Zlat (Eds.), *Encyklopedia książki. T. I, A-J*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Świgoń, M. (2018). Informatologia w analizie trendów. In A. Kucner, R. Sierocki, & P. Wasyluk (Eds.), *Trendy. Interpretacje i konfrontacje* (pp. 126–137). Olsztyn: Uniwersytet Warmińsko-Mazurski w Olsztynie.

Tabakow, M., Korczak, J., & Franczyk, B. (2014). Big Data – definicje, wyzwania i technologie informatyczne. *Informatyka Ekonomiczna, 1*(31), 138–156.

TQMSOFT, Czarski, A. (2018). *Jaką nową wartość dla firm wnosi Big Data?* Retrieved from https://tinyurl.com/y24pojed

Tuppen, S., Rose, S., & Drosopoulou, L. (2016). Library catalogue records as a research resource: introducing 'a big data history of music'. *Fontes Artis Musicae, 63*(2), 67–88.

Villanova University. (2019). *Falvey Memorial Libra*ry. Retrieved from https://library.villanova.edu/

Wójcik, M. (2016). Big data w zarządzaniu informacją – przegląd wybranych zagadnień. In S. Cisek (Ed.), *Inspiracje i innowacje: zarządzanie informacją w perspektywie bibliologii i informatologii* (pp. 61–70). Kraków: Biblioteka Jagiellońska.

Żemła, E. (2018). *Analitycy Big Data – współcześni wróżbici? Niezwykły projekt polskich naukowców*. Retrieved from https://tinyurl.com/y2j9w57r

Abstract

The modern information environment is dynamic and characterized by the speed with which multimedia content is created, collected, contributed to and shared. Users have access to documents which are part of large, changing and diverse sets of data, whose effective processing can lead, and frequently does lead, to new knowledge being discovered. However, the overwhelming majority of the resources available today require specialized tools and techniques for identifying, searching, collecting and organizing the large volumes of data. This also applies to data directly related to the activities of institutions dealing in information or the field of information science, especially the theory and practice of accessing, searching and collecting documents.

The purpose of this article is to present selected issues and challenges related to exploring the possibilities offered by big data from the perspective of information science, the activities of libraries and the information resources they offer. Based on a critical analysis of the relevant literature and with use of inductive reasoning, experiments and observations, selected aspects of digital document accessibility and classification are presented, in addition to issues related to searching and identifying resources using tools currently offered by libraries (in particular discovery systems).

Keywords: big data, data management, trend watching, information science, data sets exploration

Grzegorz Gmiterek (Doctor of Humanities in the field of book studies and information science) is Adjunct at the Faculty of Journalism, Information and Book Studies of the University of Warsaw. His research interests focus on the applications of new technologies in cultural and science institutions (in particular the tools and services of Web 2.0, as well as mobile devices and applications). He is a recipient of a scholarship granted by the Dr Maria Zdziarska-Zalezka History and Literature Association in Paris, as well as a participant in the US Department of State International Visitor Leadership Program "Library & Information Science." He is the author of several dozen publications, including the book *Biblioteka w środowisku społecznościowego Internetu. Biblioteka 2.0* (*The library in the social internet environment. Library 2.0,* 2011), for which he received the Adam Łysakowski SBP Science Award, as well as being the co-author of the book *Aplikacje mobilne nie tylko w bibliotece* (*Mobile applications not only in the library,* which received a technical and science publication award of the Rector of the Warsaw University of Technology at the ACADEMIA Academic and Science Book Fair). For the past two years, he has been working on the research grant *Exploring sources of data on the topic of R+D+I activity*, awarded by the National Center for Research and Development.