The Journal of Academic Librarianship 49 (2023) 102646

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

Big data-driven investigation into the maturity of library research data services (RDS)

Marek Nahotko *, Magdalena Zych, Aneta Januszko-Szakiel, Małgorzata Jaskowska

Jagiellonian University, Institute for Information Studies, Kraków, Poland

ARTICLE INFO

Original content: Research data set for publication "Big data-driven investigation of the maturity of library research data services (RDS)" (Original data)

Keywords:
Research data management (RDM)
Research data services (RDS)
RDS maturity
Academic libraries
Datafication
Data librarianship

ABSTRACT

Research data management (RDM) poses a significant challenge for academic organizations. The creation of library research data services (RDS) requires assessment of their maturity, i.e., the primary objective of this study. Its authors have set out to probe the nationwide level of library RDS maturity, based on the RDS maturity model, as proposed by Cox et al. (2019), while making use of natural language processing (NLP) tools, typical for big data analysis. The secondary objective consisted in determining the actual suitability of the above-referenced tools for this particular type of assessment. Web scraping, based on 72 keywords, and completed twice, allowed the authors to select from the list of 320 libraries that run RDS, i.e., 38 (2021) and 42 (2022), respectively. The content of the websites run by the academic libraries offering a scope of RDM services was then appraised in some depth. The findings allowed the authors to identify the geographical distribution of RDS (academic centers of various sizes), a scope of activities undertaken in the area of research data (divided into three clusters, i.e., compliance, stewardship, and transformation), and overall potential for their prospective enhancement. Although the present study was carried within a single country only (Poland), its protocol may easily be adapted for use in any other countries, with a view to making a viable comparison of pertinent findings.

Introduction

Data boast an increasingly significant role in creating prevalent imagery of everyday reality, especially when it comes to science. In fact, attempts have always been made to have complex socio-cultural reality effectively transposed to simple data that may subsequently easily be aggregated. This drive is clearly manifest in the emergence of a brand-new term, i.e., datafication (Mayer-Schönberger & Cukier, 2013) construed as presenting a phenomenon under study in a quantified and aggregated way, e.g., in a tabularized format, to facilitate its subsequent analysis. It is a manifestation of the times we live in, so deeply preoccupied with data processing that it makes some believe that data is the only key to actually appreciating and comprehending the reality around (Van Dijk, 2020). This is so because datafication may effectively relate to various constituent components of reality, e.g., words, geolocation, inter-relationships, experience, individual mood, bodily variables, and lots more. These data are analyzed with the aid of big data tools, specifically designed for processing large quantities of data in the manner that would otherwise pose significant logistical challenges in their manipulation and overall management (big data, 2022).

Libraries boast a place of special significance in big data processing and management ever since their computerization commenced back in the 1960s, consequently spawning the development of large, electronically structured data resources, e.g., catalogue metadata, as well as facilitating access to commercially offered bibliographic databases. As part of the technological revolution in libraries, the advent of the Internet resulted in the transfer of these structured databases to global networks, even as the library websites began to appear, each one of them offering unstructured data in various formats (e.g., text, graphics), although still regarded as quite valuable by their users. Obviously enough, the concept of what might actually be valuable for users has evolved over time. Presently, it is considered that any resources related to research data management (RDM) are popular among the regular users and much appreciated.

Both types of library data (structured and unstructured) have been used for purposes other than the originally intended ones (e.g., information retrieval). More and more often a characteristic feature of these resources consists in their use as autonomous sources, the processing of which is no longer related solely to their original procedural constraints or objectives. These data are occasionally used in quantitative research

* Corresponding author.
E-mail address: marek.nahotko@uj.edu.pl (M. Nahotko).

https://doi.org/10.1016/j.acalib.2022.102646
Received 11 August 2022; Received in revised form 5 November 2022; Accepted 16 November 2022
Available online 29 November 2022

in quite surprising ways, in a different place and time, potentially by anyone with access to the new communication technologies infrastructure. Bibliographic and catalogue data, assembled and made available in the libraries, have long been the focus of academicians looking for the resources for quantitative research which makes use of the big data techniques. In recent years, bibliographic data science has been proposed as a new approach to bibliographic data, particularly catalogue metadata (Lahti et al., 2019).

It has been noted that a systematic analysis of library catalogues yields a wealth of information on the historical dynamics of knowledge production and dissemination (Tolonen et al., 2019). In bibliographic data science, library metadata is treated as a research object based on more general open science and data science paradigms, also treated as open data science (Uhlir & Schröder, 2007). Data mining in relation to library data is also construed as bibliomining (Tu et al., 2021, 3), i.e., the study and mapping of user behavior and resource use. This practice facilitates collecting unstructured data, as they are being generated by the librarians during routine library management and service delivery processes, or by the library users themselves (Liu & Shen, 2018), unlike the bibliographic (structured) data generated through cataloguing.

This study addresses three closely related research areas. Firstly, we are going to focus on these unstructured library data which are less frequently analyzed, but, when analyzed, may offer some valuable insights as to how overall user satisfaction level might be boosted, and the value of information services at large (Ahmed & Ameen, 2017). The overall aim consists in presenting the potential for making use of unstructured data and information resources collected on the websites of academic libraries as the data sources for quantitative scientific research. By way of gaining insights into the RDM processes of academic libraries, as available on their websites, the maturity of the RDS provided by these libraries was assessed. Academic libraries usually offer on their websites a variety of information about managing research data at different stages of their lifecycle (Lewis, 2010). Assessment of the actual content of the library websites pertaining specifically to research data allowed the article authors to determine overall maturity level of their services.

Secondly, the analyses of unstructured data were carried out, while making use of the techniques typical for big data and natural language processing (NLP) which allows the use of data and information resources created to support library users for other purposes, previously not anticipated by the creators of these data and information resources. The big data techniques applied, e.g., web scraping, make it possible to use the information resources in a similar research not only with regard to the libraries, but also to other GLAM sector institutions (i.e., galleries, libraries, archives, and museums), as well as other institutions providing a scope of cultural/intellectual services. Our research demonstrated that library websites contain resources that may be used as an indicator of their datafication level. These data can be managed systematically, taking due account of local specifics.

Thirdly, the study addresses the maturity of library services related to research data (RDS) in the area through the assessment of the library resources related to research data management (RDM). This means that the research data are treated in the article on a multi-level basis (Fig. 1). On the one hand, data librarians use the tools such as the library websites to post their information resources relevant to the information needs of their users. In this case, there were the academic investigators who create and subsequently use as well as reuse the research data in the research cycle. In this way, the librarians run data management and a scope of data curation activities which are collectively termed as data librarianship (Semeler et al., 2019). On the other hand, analysis of these resources, aided by big data tools, may identify their specific properties which are in fact indicative of the level of RDS maturity. It is therefore possible to determine the current status and development directions of the library's RDS, its benefits for the users, and key conditions for further development, as this actually pertains to the individual level of professional expertise to be commanded by the librarians.

The maturity of RDS is construed as being in continuous evolution, i.e., evolving from primary to highly advanced level (Tiwari & Madalli, 2021), with principal focus resting on professional excellence. While maturity might imply the completeness of these services, this is not the desired status to be actually achieved, but rather to be pursued as an evolutionary progress towards a specific objective. The maturity assessment is one of the conventional approaches used for determining the level of sophistication in the range of the services rendered (Kouper et al., 2017). The maturity model is a tool used for assessing the current status and any future prospects for any attendant process, person, or a group of individuals involved in RDS. It is also used to identify the specific factors that might prove instrumental in a prospective enhancement of this service.

Even though the actual scope of the present study is confined to the library environment within a single country (Poland), the study protocol is easily adaptable to pursue similar, comparative studies in other countries. Any internationally pursued studies stand a good chance of providing much broader insights into the select processes, prevalent trends, and the actual impact of a specific geographical location on RDS. This in turn might well offer valuable pointers as to what might be anticipated in that regard in the foreseeable future.

Theoretical background: related literature

Currently, intensive research is carried out in the areas referenced in the Introduction. A number of earlier published studies are reviewed further below, divided into four key areas.

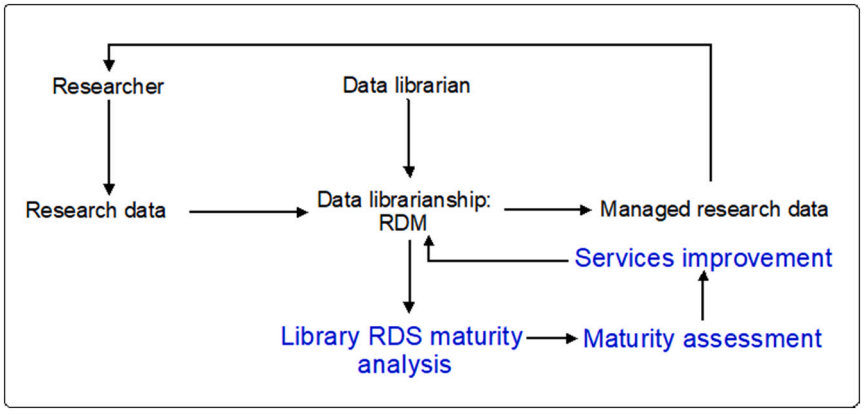

Fig. 1. The positioning of RDS maturity investigations into data librarianship.

Datafication, data science and data scientists

The rapid development of data recording, processing, and storage technologies resulted in a dynamic increase in overall volume of assembled research data. Such growth dynamics pose a challenge in their routine use and reuse. Extracting specific body of knowledge from large collections of research data requires their prior reduction and aggregation, i.e., a process termed datafication (Mayer-Schönberger & Cukier, 2013). This entails presenting the phenomena under study in a quantified form, and in an aggregated manner, thus enabling their further processing.

The greater frequency of large data sets processing, along with rapid development of the methods and techniques of such processing have brought about manifest changes in the perception of body of knowledge being extracted in this way, as well as of the extraction process itself. Such trends are well exemplified by the data-centric way of perceiving reality, effectively shaping one's mindset, as well as setting one's mental processing algorithms into the "higher authority" mode (Osika, 2021). Entrenching oneself in such an assertion may well lead to "data-ism", that is, an actual belief in the limitless potential of algorithms as effective tools for reading the "world paradigms", as well as deeming the information inflow the absolute value within its own right (Brooks, 2013; Harari, 2017, 75).

As the key definitions in this field have not been standardized as yet, this has given rise to numerous terminological considerations regarding the processes of transforming the pertinent resources into a digital form, as well as the tools for the "objectification of knowledge-forming processes" (Goczyła, 2021). For the purpose of this study, the authors use the term 'datafication', construed as the process in which the social activities, their agents, objects, and routine practices are transformed into quantified digital data (Mayer-Schönberger & Cukier, 2013). In this process, components of reality, from physical quantities through socio-demographic properties, to a body of human experience and emotions, are subject to the processes of registration, quantification, aggregation, and algorithmisation, with a view to finding their equivalent in the data stream (Iwasiński, 2020).

The value of research data, especially those generated in the course of publicly financed research, stems from the need and potential for their systematic sharing and reuse (Feger et al., 2020; Karasti et al., 2006). Open access to these resources and their dissemination boasts numerous advantages, as compared to a closed system which significantly hinders access and any subsequent reuse of such data. It aids maximizing the research potential of new digital technologies and communication networks, thus boosting the revenues generated from public research investment (Arzberger et al., 2004).

An increased volume of data in digital format has triggered the need for pursuing formal research into the area. A diversity of phenomena occurring throughout the "research data lifecycle" have become the subject of academic inquiries, investigations, debates, as well as pose numerous research questions and hypotheses. This has ultimately prompted the emergence of the field of science known as data science (also called e-science), defined as "an interdisciplinary field using scientific methods, processes, algorithms, and systems to extract insights from many structured and unstructured data" or, shorter still, "the study of the generalizable extraction of knowledge from data" (Dhar, 2013). Data science is related to data mining, machine learning, and big data analysis, and it uses statistics, data analysis, machine learning, expertise in specific domains, and related methods, in order to comprehend, appreciate, and analyze data (Hayashi, 1998).

The authorship of the term *data science* is attributed to Peter Naur, Danish IT pioneer and Turing Award winner. The researcher was probably the first one to use this term in 1960. He proposed that data science was the study of the generalizable extraction of knowledge from data (Naur, 1974; Wainer, 2015, 2; Virkus & Garoufallou, 2019). Further uses of the term emerged during the 1960s and 70s, construing it as the science of dealing with data, once they have been established, whereas the actual relation of data to what they represent is delegated to other fields of knowledge and academic domains (Naur, 1974).

The review of literature on the subject attests to the fact that the term *data science* is defined in many ways, even though two distinct approaches may readily be identified:

- a general, chronologically older approach, in line with which the data science, apart from dealing with the processing of large data sets (in order to extract knowledge), also covers the issues related to the collection, development, management, and storage of large data sets (Ratner, 2017; Stanton et al., 2012),
- a narrower, currently dominant approach proposes reducing the data science down to the techniques, methods, and tools used to obtain specific values and insights from the data sets (Provost & Fawcett, 2013; Song & Zhu, 2016; Amirian et al., 2017; Kelleher & Tierney, 2018).

Along with defining the meaning of the term *data science*, there is an ongoing debate on the legitimacy of establishing it as fully autonomous, brand-new academic discipline, instead of regarding it as a mere extension of statistical methods (Cleveland, 2001; Diggle, 2015). Some researchers also associate data science with such terms as business analytics, operations research, business intelligence, competitive intelligence, data analysis and modelling, and knowledge extraction from big data (Foreman, 2013; Kelleher & Tierney, 2018).

Regardless of the arguments related to data science terminology, the literature on the subject highlights the importance of specialist (professional/domain) expertise, which tangibly aids understanding and appreciation of the true problems and realities of specific fields of knowledge. This way the actual application of data science in resolving true-to-life challenges in particular academic domains is put into its proper context (Sanchez-Pinto et al., 2018). Recent years have brought more widespread application of data science in many fields - from business and economics, through public policy, to science and education (Voulgaris, 2014). The acquired body of knowledge can help describe what happened, explain why something happened, and predict what might happen (Baškarada & Koronios, 2017), contributing to the optimization and increase in the efficiency of the processes (Granville, 2014).

Indubitably, there is a growing demand for staff boasting sufficient expertise in broadly construed data resource management, focused predominantly on data analytical skills. Data analytics has even been described as the most attractive (sexiest) job of the 21st century (Davenport & Patil, 2012, 70). No clear definition of either a data scientist, or a data manager has as yet been established and adopted. At the moment, these professionals are known by a variety of titles, e.g., data steward, data analyst, data engineer, data curator, depending on the specific stages of data management and handling within the data lifecycle they happen to be involved in. Overall expertise combined with a multitude and sheer diversity of tasks that actually make up the profession of an academic researcher make it downright impossible to be found in a single individual. No one can be an expert in so many areas at the same time. This merely goes to show there is a need to build the teams of individuals well adept in many different areas of expertise required for RDM tasks (Schutt and O'Neil, 2014, 10).

Datafication, a single aspect of this broad area of expert knowledge, is being investigated in terms RDM and the attendant services which support the processes (RDS). It is construed as an area of interest and an essential practical activity of academic libraries which have taken up overall responsibility for building the research data repositories and supporting both the scientists pursuing their research, and the university administrators in their management of research processes (Laskowski, 2021).

Data librarianship and research data management (RDM)

Data science may well be deemed a new area of research for IT professionals and librarians, otherwise called data librarians (Semeler et al., 2019) aiming to address the issues related to data management and analysis. Data librarians focus on disseminating important research findings in the form of relevant information by collecting, organizing, and managing data from multiple sources. They function as the facilitators of research at all stages of scholarly research cycle, by providing all kinds of data, information, and services required in the data management and curation process. Hence, their scope of activities is all the more closely tailored to the requirements and tasks pursued by the researchers within the scope of RDM (Al-Jaradat, 2021). At the same time, the data librarian becomes a data scientist who uses his expertise on data and computing to create brand-new, data-driven products and services that support new forms of data-intensive scholarship (Tang & Hu, 2019). The field of data librarianship constantly and systematically probes the competencies and responsibilities of its professionals in terms of current and future e-science trends. To take this even further, some professionals believe that a data librarian should possess the same standard of skills as a data scientist (Xiu & Wang, 2014).

Currently, data librarianship is focused on creating new library services based on new approaches to supporting RDM and curating digital data created during scientific research (Koltay, 2017; Singh et al., 2022). One of the stages of RDM maturity development is the application of the principles, practices, and resources of traditional librarianship to research data. It is, therefore, posited that data librarians should have the technical skills and knowledge necessary for data management, data curation, and competences to support RDM as the necessary basis for the provision of RDS supporting data science and scientific communication (Tenopir et al., 2014). The expectation is that data librarians should be able to collaborate with scientists carrying out research during all stages of their work in the documentation processes, ensuring that these operate effectively (Xu et al., 2022). McCaffrey and Giesbrecht (2016) presented several skills and competencies of data librarians, identifying three main areas of activity for this profession: a) data management and curation; b) data visualization and geospatial representation; and c) advanced information services.

The qualifications and competencies of the data librarian is likely to affect the quality of the RDM services, the immaturity of which researchers have pointed out. For example, researchers have observed the immaturity of RDM services in libraries, such as the lack of appropriate union catalogues (Yoon, 2017). A 2019 study indicated that 35 % of libraries in the US offered data curation services, and a further 15 % were preparing to implement them (Yoon & Donaldson, 2019). Data management planning (DMP) and data organization/description were the two most frequently proposed services while later studies indicated training and data storage in repositories as being a priority (Yidavalapati et al., 2021).

According to Whyte and Tedds (2011), RDM is concerned with the organization of data from its entry into the research cycle to the dissemination and archiving of valuable results. It aims to effectively verify the results and allows for new and innovative research based on existing information. RDM is usually associated with the preservation and curation processes, the former dealing with the provision of long-term retention of data in repositories, or at a data journal publisher for reuse. The latter ensures that the results of a research project can be archived and reused many times. Issues of data quality are also important (Al-Jaradat, 2021).

Another definition states that library RDM encompasses many activities and processes related to the research data lifecycle, including data design and creation, storage, protection, preservation, retrieval, dissemination, and reuse, considering technical feasibility, ethical and legal considerations, and organizational framework (Cox & Pinfield, 2014). Similarly, the activities of RDM are presented by the OCLC, proposing three categories of these activities: 1) educational services to highlight the importance of RDM, the training of researchers in basic RDM skills, and guidance to RDM tools and resources; 2) expert services, offering decision support and customized solutions to RDM problems indicated by the researchers; 3) curation services for technical infrastructure and related services that support RDM throughout the research process (Bryant et al., 2017).

Kim and Syn (2021) distinguished six stages in the data lifecycle: 1) creating new data or reusing existing data; 2) description (metadata); 3) collecting data in an appropriate format; 4) data analysis to achieve results; 5) dissemination in data journals, repositories, etc.; 6) long-term archiving. The stages were mapped into three types of library services: education, expertise, and curation. Yoon and Schultz (2017) surveyed 185 websites of US academic libraries and categorized their research data resources into four main areas of interest, i.e., service, information, education, and network. In the conclusions, they asserted that libraries should be more involved in the delivery of RDS, the provision of online information, and the creation of educational services. Of particular note was the fact that there were no services related to the iterative use of data.

Kouper et al. (2017) proposed a typology of RDS implemented in libraries. Based on the literature and their own research, they distinguished three groups of services, which allows for uniform comparison of services between institutions and determining their maturity. This breakdown is as follows:

- core group: Data management plan (DMP), assistance and mandate support, consultations, and instructions, best practices,
- intermediate group: data deposit and repositories, archiving and preservation, metadata, storage, sharing and reuse,
- advanced group: data processing and analysis, data curation, acquisition and citation, copyright, software and hardware, policies, data reference.

The participants in the RDM processes comprise scientists, data managers, IT professionals, librarians, and archivists, among others (Whyte, 2014). For this reason, satisfying the diverse needs of all the participants involved in the data lifecycle may not be assigned to a single department within any organization. RDM, as interdisciplinary specialty operates under the conditions of mixing knowledge and skills (Ran et al., 2021). Therefore, collaboration and partnership between the library, the IT department, and other internal and external units of the organization are essential for the development and success of RDM services (Delserone, 2008; Pinfield et al., 2014; Verbaan & Cox, 2014).

Cox et al. (2019) believe that RDM services could bring about a fundamental change in the role played by academic libraries in the institution due to their increasing involvement in supporting scientific research through deeper involvement in the research process, even to the librarians participating directly as a member of the research team. Another important aspect is the decline in the role of the library as providers of literature from outside the organization and a shift towards the internal production of data, information, and knowledge which brings more focus on curation, preservation, and reuse. This, in turn, demands new competencies among library professionals especially with regard to activities such as data analysis, visualization, and research integration, thus, making them data librarians (Berman, 2017).

Despite discrepant concepts as to what exactly constitutes library RDS, there are some commonly identified features of these services. One is the consulting services, seen as having a distinct advantage over the many practical or general technological activities carried out by the libraries (Tenopir et al., 2017). This may be related to the practice of matching RDS to the scope of services traditionally provided by the libraries. Another common phenomenon is the lack of RDM competency among the library staff (Xu et al., 2022). In general, librarians show a strong commitment to RDM activities, although to a different extent, depending on the region and the actual type of library.

Big data in libraries

Understanding big data is often intuitive (Garoufallou & Gaitanou, 2021, 411; Li et al., 2017). Laney (2001) presented the characteristics of big data as data that cannot be processed by traditional data management tools. Others (Dumbill, 2013) suggested a definition of big data as data that exceed the processing capacity of conventional database system, calls for finding alternative ways of processing them. Big data are data generated constantly, automatically, and rapidly (Reinhalter & Wittman, 2014). Business leaders from around the world have for years recognized the importance and value of big data analytics due to its enormous operational and strategic potential, while generating a demand for data specialists (Manyika et al., 2011; Provost & Fawcett, 2013).

In academic libraries there is a large amount of data, both structured and unstructured (Tella & Kadri, 2021), covering the entire spectrum, from structured metadata sets, e. g., Online Public Access Catalog (OPAC) data, electronic content in various forms, and large amounts of unstructured data of all kinds, including those made available on the Internet. These data are supplemented with constantly growing resources of information from various sources available on the Web, commercial and free. Collectively, academic libraries are now having to deal with a variety of data resources in the form of not only traditional books, journals, and bibliographic data but also all other forms of data such as textual documents, metadata, images, audio and video files, research data, software, and 3D collections. This means the library involvement with big data (Garoufallou & Gaitanou, 2021).

The source of unstructured data is mainly the content of library websites and social media sites, such as Facebook and Twitter. As reported by Gardner et al. (2008), websites are increasingly used by academic libraries to promote their services and resources to users-researchers. These websites contain unstructured content on matters, such as library collections (collection management, scientific communication, collection policy, links to commercial databases) and library services (e.g., service for unregistered users, information about ILL, circulation, reference, and reservations). Gardner et al. (2008, 19) distinguished such categories of content posted on library websites as those about the library, current problems (including scientific communication, open access, digital library, institutional repository), collections (electronic collections, catalogues/databases, citation management software, new acquisitions, special and multimedia collections), services (loans, acquisition, research support, teaching support), and contact details. Pareek and Gupta (2012) indicated such information about the services of Indian libraries as loans, Internet access, information services, reprographic services, and bibliographic services. Manjunatha (2016), while examining the websites of eight libraries, distinguished such categories of information as general information about the library, information about the library's collections, information about library services, the availability of electronic resources, and links.

According to Ball (2019), there are three ways to apply big data in academic libraries, namely: data sources, data analysis (library analysis) and data visualization. These uses of big data by librarians can lead to overall improvement of library services based on a better understanding of the needs of users and their information seeking behavior (Provost & Fawcett, 2013). Planning services based on this kind of knowledge should result in increased efficiency of library services, greater resistance to the hacker actions, and the search possibilities boost (Hoy, 2014; Tella & Kadri, 2021; Zhan & Widén, 2019). In this approach, the library is a place where it is possible to use, store, and organize data, including big data.

Our idea is to use these (unstructured) data for analysis that goes beyond traditional library activities. It concerns the study of phenomena that are more general than those directly related to the information activity of libraries. In this case, library data are used there to detect phenomena occurring outside the library, as in the above-referenced bibliographic data science, where linguistic and literary phenomena are studied using metadata analysis. Further on in the study, we will present an example of such research dealing with the state of datafication of science in Poland on the basis of library data.

Library RDS maturity

In this study, unstructured big data from the libraries were used to study the maturity of information systems with particular reference to RDS. Maturity assessment is a frequently implemented approach to determine the level of sophistication of a process, service, or a product (Kouper et al., 2017, 159). The concept of maturity is usually construed very broadly as something fully developed or perfect (Cooke-Davies, 2004, 1234). Maturity models have been developed that describe the improvement process and set the directions for development, including the criteria and indicators defining the existing and desired state. The ideas of Total Quality Management (TQM) are often indicated as their source, which is confirmed by the early work of Crosby (1979), who presented the Quality Management Maturity Grid, containing five levels of maturity moving from ad hoc tasks to a fully implementable and systematic approach.

Since the end of the last century, maturity models have been applied to evaluate the development of software systems. One of the early, better-known examples is the Capability Maturity Model for Software (CMM-SW), which was developed in the 1990s at Carnegie Mellon University to facilitate the US government's procurement of appropriate software (Yang et al., 2016). The purpose of this model was to evaluate software processes in order to move organizations away from a chaotic and unplanned development process towards a more disciplined and optimized one. The creators of the model sought to distinguish between immature and mature programming organizations by arguing that the former was focused on reacting and solving currently emerging problems on an ongoing basis, while the latter was based on sound management techniques.

This model has been used in work aimed at assessing and improving RDM activities in research projects (Qin et al., 2015; Qin et al., 2017). Following the earlier model of Humphrey's (1989) CMM (Capability Maturity Model), they presented five levels of maturity in relation to RDM:

- RDM entry level, based on the competences of individuals and their individual efforts, making the work unreliable;
- The management level, based on procedures and policies individually developed for each project, which makes it difficult to adopt solutions between projects;
- Defined level, characterized by accepted and repeatable procedures that can be used in several projects;
- Quantitatively manageable level that allows the use of measures that facilitate the evaluation of processes and progress;
- Optimal level at which weaknesses and inefficiencies are identified and then actively removed.

The maturity levels are determined in the model for five main processes and areas of practice: 1) data management in general; 2) data collection, processing, and quality assurance; 3) description and representation of data; 4) data dissemination; 5) repository services and preservation. For each area of practice, there are also pertinent reference values for its evaluation.

A more empirical approach was proposed by Kouper et al. (2017), who presented only three levels of maturity: basic (creating foundations), intermediate (organization and standardization) and advanced (monitoring and optimization), but as many as eight areas of research: 1) leadership (vision, strategy, culture); 2) services; 3) users and stakeholders; 4) supporting the research life cycle; 5) management; 6) costs and budgeting; 7) cooperation between units; 8) human capital. The model creates a matrix with areas in the rows and maturity levels in the

columns, and it has two goals - identifying weaknesses and setting priorities, as well as functioning as a communication tool with the library administration. According to these authors, mature RDS programs offer activities best suited to the mission of the library and its parent institution.

Cox et al. in 2017 proposed a four-tier RDM maturity model, which was modified in 2019. In this new version of the model, more attention was paid to activities such as data analysis and visualization, as well as their integration (Table 1).

These levels are categorized according to the existence or absence of services and support, compliance, skills, roles and structures, practices, and cultural acceptance.

Methodology of RDS maturity investigations

Previous studies of library electronic resources conducted outside of libraries were concerned with collected structured data, mainly descriptive metadata (Lahti et al., 2019) or unstructured data, such as digitized texts of publications from library collections (Leetaru, 2015). The study of the content of library websites was of an auxiliary nature, as a supplement to methods such as surveys and interviews (Singh et al., 2022, 3). So far, there has been no research on the websites of academic libraries (unstructured data) as a source of information on trends in RDS. This approach is represented in the studies presented in this section and is based on the findings from the hypothesis that many phenomena that function in science (including research data processes, such as RDM), are reflected in the activities of academic libraries via their websites.

Due to the relationship between data librarianship and data science, it is possible to study the level of datafication of scientific organizations with the use of big data methods. Data librarianship focuses on creating new library services – RDS, and transforming existing scientific consulting services based on new ways of managing and curating digital research data. Therefore, given the big data methods, especially in combination with NLP tasks (Gudivada et al., 2015) used in data librarianship, it is possible to determine the degree of maturity of these activities which may help improve the practical effects of data librarianship. Therefore, this is an area that deals with the results of the work of data librarians and big data scientists.

Applying data librarianship to academic library data allows the conduct of science mapping analysis, which is a set of methods and techniques designed for creating science maps (Petrovich, 2021). These methods help to find answers to questions such as:

- What is the level of library RDS maturity and how is it different in terms of the type of library?
- In which disciplines is RDM better and in which less developed?
- In which research centers are research data issues of interest?
- What are the relationships between respective research data problems?

The answers to these questions help in understanding the ways in which the structural units of science are interrelated at different levels (Leydesdorff, 1987).

Data sources: web pages of academic libraries (unstructured data)

The investigative team opted to include the websites of all academic libraries in Poland in the first phase of the study. For this purpose, the existing list of web addresses of these libraries available on the librarians' portal called EBIB (http://www.ebib.pl/biblioteki/baza/) was used. The list is constantly updated, which increases its pertinence. It was necessary to analyze the source code of the EBIB web page, as it was a frame structure. Thanks to this, access to the html source code was obtained, which allowed to save the list of libraries as an html file. The elements that originally facilitated the use of the list on the EBIB website, but made it difficult to perform web scraping have been removed from this list.

From this list, the web addresses of 320 academic libraries were taken. Then, those that offer RDS and provide information on this topic on their websites had to be selected.

Data analysis: web scraping

In the next step, based on the list of addresses taken from EBIB, websites were selected based on their content using the web scraping technique. Web scraping is automated data extraction from the Internet whereby it is possible to obtain structured data as a result of unstructured textual data transformation (see Arbia & Nardelli, 2020; Lunn et al., 2020; Regueira et al., 2020; Uzun, 2020, 61726). As explained by Yan (2020) and Dongo et al. (2020), as well as others, web scraping is particularly suitable for processing large data sets because of the automation of repetitive activities, which allows for quick and accurate data acquisition from the Internet.

The web scraping method was used in the study to obtain a list of web addresses of Polish academic libraries from the publicly available EBIB list and to search the obtained web addresses for the presence of keywords in the content of their source code. A list of 72 keywords containing terms related to RDM issues was prepared (see Annex 1). This list is based on the expertise of the members of the research team. These keywords have been prepared in such a way as to take into account the inflection of the Polish language. This list enabled the selection of web addresses for further evaluation using big data tools for data analysis and visualization.

The adopted web scraping procedure consisted of several stages and included analysis of the source code of the web pages, searching the pages with their internal subpages according to the list of keywords, and downloading and saving the results to files in CSV format. The script code for web scraping was written in the Google Colab environment using the Python programming language and its selected libraries, i.e., BeautifulSoup,[1] requests[2] and urllib.[3] In addition, libraries for data analysis and processing were also used: CSV[4] and pandas[5] (for saving data to result files) and re[6] (for natural language processing in terms of the conjugation of keywords in accordance with the rules applied in the Polish language). Then, the resulting CSV files were then processed into MS Excel spreadsheets for later analysis.

The web scraping procedure with the use of a prepared list of keywords was performed several times, in mid-2021 and 2022, which made it possible to determine the dynamics of the changes in the number and content of selected websites. The need to repeat the research resulted from the belief that the situation in the library RDS area in Poland is very dynamic. This way of operating allowed us to capture these dynamics. In 2021, 53 library websites were found where keywords from the list were detected, but further analysis resulted in the rejection of a part of them, so only 38 libraries remained. This resulted from the unsuitability of the content of some of the library websites, where, for example, there was only a link from the branch library page to the main library's RDM website. In this case, only the main library website content was included, and the subpage content was discarded as unsuitable for further analysis. In 2022, there were 42 such libraries, in that six new libraries appeared, including libraries from two large classical universities and two medical universities. All websites were extracted as a result of them being full-text searched (including their internal subpages) in terms of the incidence of words from the original list of keywords in their source code. For the list included 72 keywords returns

[1] https://pypi.org/project/beautifulsoup4/.
[2] https://docs.python-requests.org/en/latest/.
[3] https://docs.python.org/pl/3.8/library/urllib.html.
[4] https://docs.python.org/3/library/csv.html.
[5] https://pandas.pydata.org/docs/.
[6] https://docs.python.org/3/library/re.html.

Table 1
Cox et al. (2019) model of evolving maturity of the RDM

Levels	Skills	Activities
0 none	Existing skills	Audit, survey
1 compliance	Translation of existing skills	RDM training, data literacy, DMP, publication, citation, storage, rights, policy
2 stewardship	Reskilling of existing staff	Data repository, selection, catalogue, curation, preservation, metadata
3 transformation	New skills acquisition	Data analysis, visualization, integrity

were obtained for 26 keywords, which means that 46 keywords from the list were not found on the library websites. The text content of the library websites selected in both years was then analyzed using big data tools.

Only RDM-related web content was submitted for further analysis. For this purpose, the content of the websites has been reformatted to plain text. The scraped websites varied greatly in size. After being translated into English, which was necessary due to the requirements of the analytical tools applied (see the part entitled Data visualization), they counted from 48 to almost 22,000 characters without spaces (5010 characters on average). The total number of characters in the sample was 185,371. The total number of words was 33,289.

The translation of the content of the selected websites into English was carried out in two steps. First, the text was translated using Google Translate. Then the result was analyzed by researchers in order to remove any errors that hindered further analysis using NLP tools. There were not too many translation errors, although the principal focus rested on correcting the actual mapping of terms (what Google Translate is useful for), whereas any stylistic deficiencies were of considerably lesser significance for the task at hand.

Data visualization

The previous stages of work allowed the isolation of selected keywords and marking their popularity on the websites of respective libraries, considering changes in the degree of maturity of their RDS over time. Before doing a deeper analysis, however, we tried to find answers to what these keywords could indicate about the geographical and institutional distribution of the phenomena they represented. The presentation of all locations obtained with the web scraping technique on the map allowed us to go beyond single keywords, websites, and services, and to view all this information resource and discover what it said about the part of the world under study.

Fig. 2 indicates on the Poland's map the spatial distribution of academic libraries selected by web scraping. The Tableau Desktop tool was used, which allows for the creation of interactive visualizations. A bar graph was used for each academic center, indicating the ID count values. The ID count used is taking into account the diversity of the intensity and maturity of RDS, which indicates the "density" of research data issues on the websites of the selected libraries. The ID count is the indicator of the "saturation" of the website with relevant keywords, taking into account the depth of the hierarchical structure of the websites. For each library,

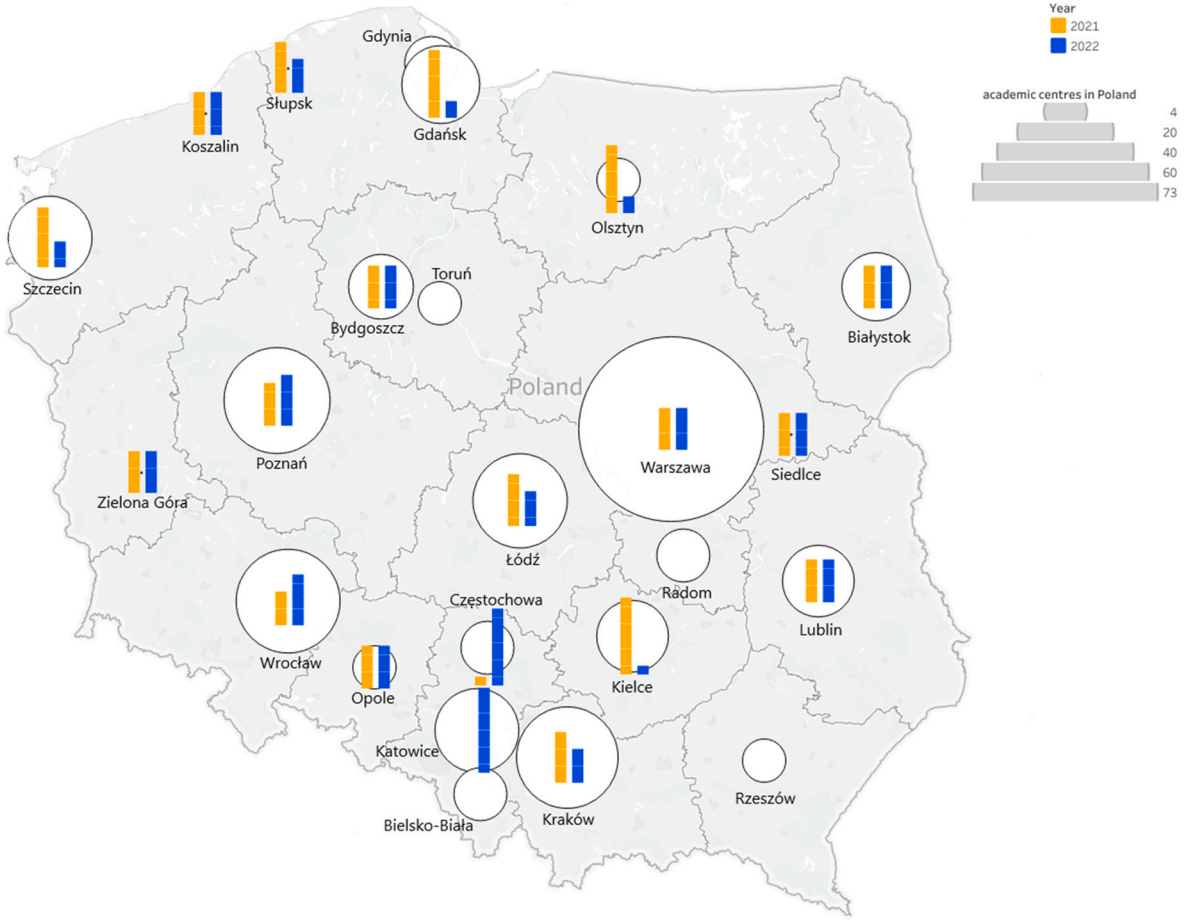

Fig. 2. Geographic visualization of data management.

in a given year, all combinations of pages and keywords in the results are counted. The ID count can then be represented as the product of two values: the keyword hits and the number of pages where the keywords were found. More precisely: ID count ∈ <MAX(keywords hit, pages hit); keywords hit x (pages hit+1)>. At a minimum, the ID count takes the value of the greater number of pairs: keywords hit and pages hit.

Fig. 2 indicates that the distribution of libraries offering RDS and the intensity of the materials they offered did not change rapidly. The places where these services were offered were the largest academic centers in Poland (Warsaw, Kraków, Wrocław, Poznań, Gdańsk, Lublin), as evidenced by the comparison of the bar graphs in Fig. 2, showing the ID count for academic center in two years and with the pie charts showing the size of the academic center. This size is determined by the number of universities operating in a specific geographic area. A dot instead of a circle means the number of universities below four. Subsequent libraries that offered new websites for research data were usually located in the same places (academic centers), strengthening them in terms of RDS offered to users. It is similar in other countries, e.g., in the USA, as indicated by research by Tenopir et al. (2015). However, in Poland, in smaller academic centers, there were also library RDS with a correlation coefficient between the size of the academic center and that of the RDS being 0.52 (moderate correlation).

Fig. 3 indicates the distribution of libraries according to their ID count. We have organized library websites along the axis from large to small in terms of their volume and keywords retrieved. The large services, containing tens of thousands of words and many keywords, are on the left-hand side of the axis, with smaller web sites sorted by decreasing size trailing off to the right. The main area under the right-hand side of the curve is the long tail of library services. These services are more difficult to find and therefore less frequently used. The long tail metaphor comes from commercial activities and serves to represent the distribution of popularity of electronic information objects (Anderson, 2004). This power law distribution is also used to describe scientific activity by size of project and scale of data production (Borgman et al., 2016). This is because the long tail can be generalized into all technology-induced areas, and science certainly belongs in there.

The situation described in Fig. 3 is also in line with the Pareto 20/80 principle, as approximately 20 % of libraries offer users 80 % of resources on research data. However, these small local sites (located in the tail) can have a significant impact on the maturity level of RDS as they offer services that target the local needs of the researchers they serve. These services may, for example, include the provision of so-called dark data (Heidorn, 2008). On the other hand, attention should be paid to the frequent need to improve the maturity of the services in the tail. In general, it can be said that the development trend of the library RDS is correct: apart from a few large services that support the big science, many smaller ones for small science are created.

Fig. 4 indicates averaged ID count values with a breakdown into libraries of various types of universities. They show the average size of the research data websites, which directly indicated their maturity and their usefulness for users. Lower values in 2022 for universities (classical - 14/16 libraries) and technical universities (13 libraries in both years) might have resulted from changes in these services. In 2021, this was content scattered across many different library websites, where RDM pages were often combined with open access pages or other materials. From 2022, libraries created more compact websites, separate from other content, usually consisting of a subpage devoted only to research data management services. The very high results for economic/business universities were due to their small number, with only two libraries from this group offering websites for research data, and one of them was very extensive.

Using analytical tools, attempts were also made to analyze the content of library sites in the direction of their maturity and usefulness to library patrons. We used free tools available on the Internet, the use of which turned out to be very intuitive. The results of these analyses, aided by VOSviewer program are presented in Figs. 5 and 6. The first visualizes the co-occurrence matrix for the selected term ("data management"), which allows for an intuitive presentation of terms related to research data. To achieve this effect, it was necessary to prepare the so-called corpus file. A corpus file is a text file that contains on each line the text of a document. The text of a document must be in English, since the natural language processing algorithms used by VOSviewer do not support other languages. The size of the nodes in network presented in Fig. 5 indicates the frequency of phrases. The thickness of the lines is proportional to the proximity of the phrases.

Using the VOSviewer program, three clusters (color-coded in Fig. 5) were selected from the content of the research data websites, containing a total of 70 phrases, the so-called "items", connected together within a cluster and between clusters. Items are represented by their labels and by default also by a bullet. The size of the label and the bullet of an item is determined by the weight of the item. The first of these clusters

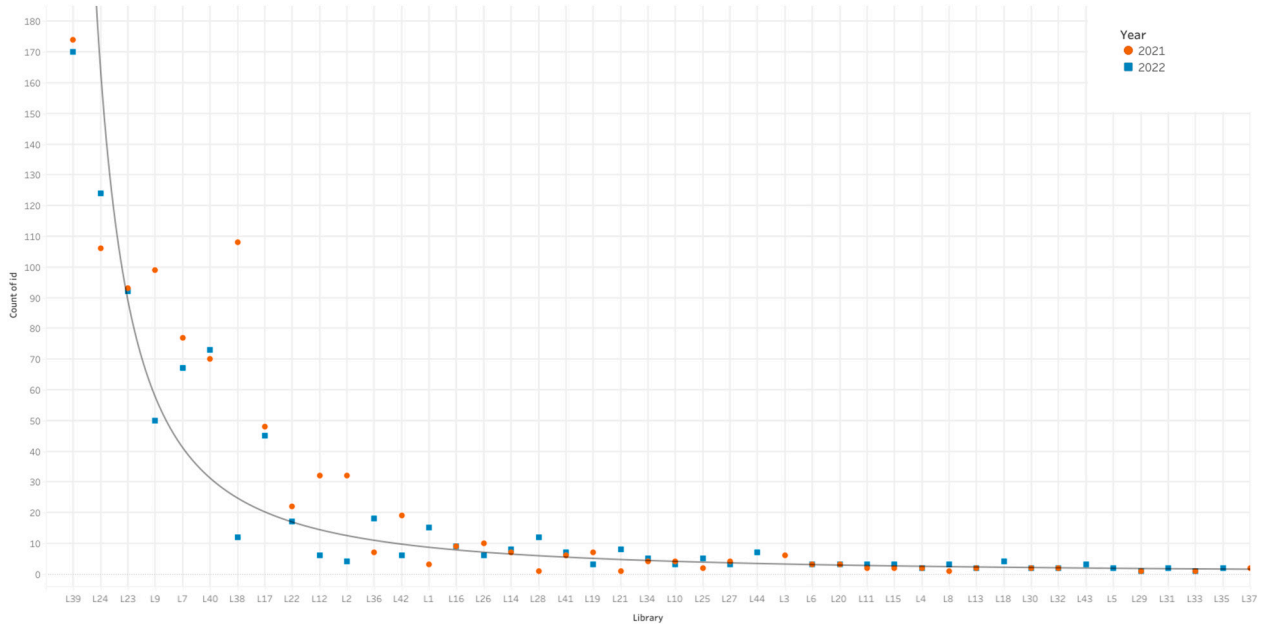

Fig. 3. Distribution of libraries by their ID count: the long tail.

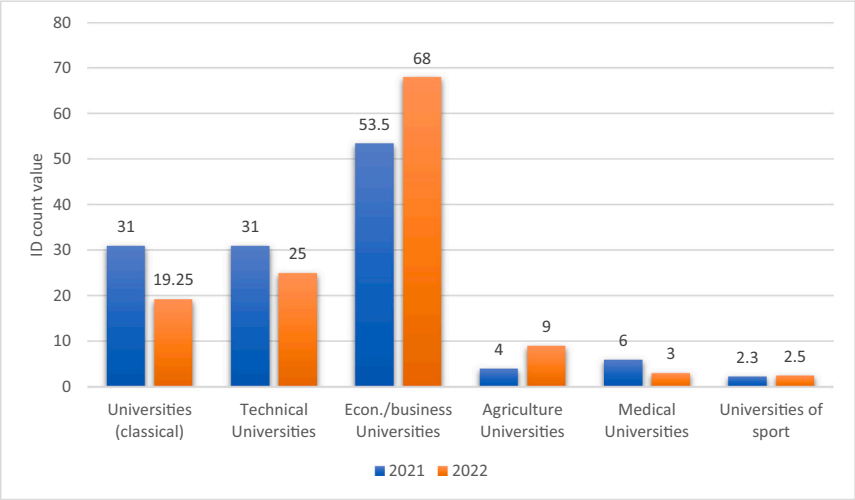

Fig. 4. ID count by types of university (averaged results).

Fig. 5. Clusters and network visualization of co-occurrence matrix.

marked in red refers to the creation of research data and their use in research (knowledge-generating activities). The second, in green, indicates the materialization of data in various forms, and the last, in blue, represents the tools used in RDM and research data applications.

To obtain more information about selected phrases, a graph of visualization of the density of phrases (keywords) was created (Fig. 6). In this visualization, each point on the map has a color that depends on the phrase density at that point. By default, this color goes from blue to green to yellow. The greater the number of phrases next to each other in the point and the higher the weight of these phrases, the yellower it

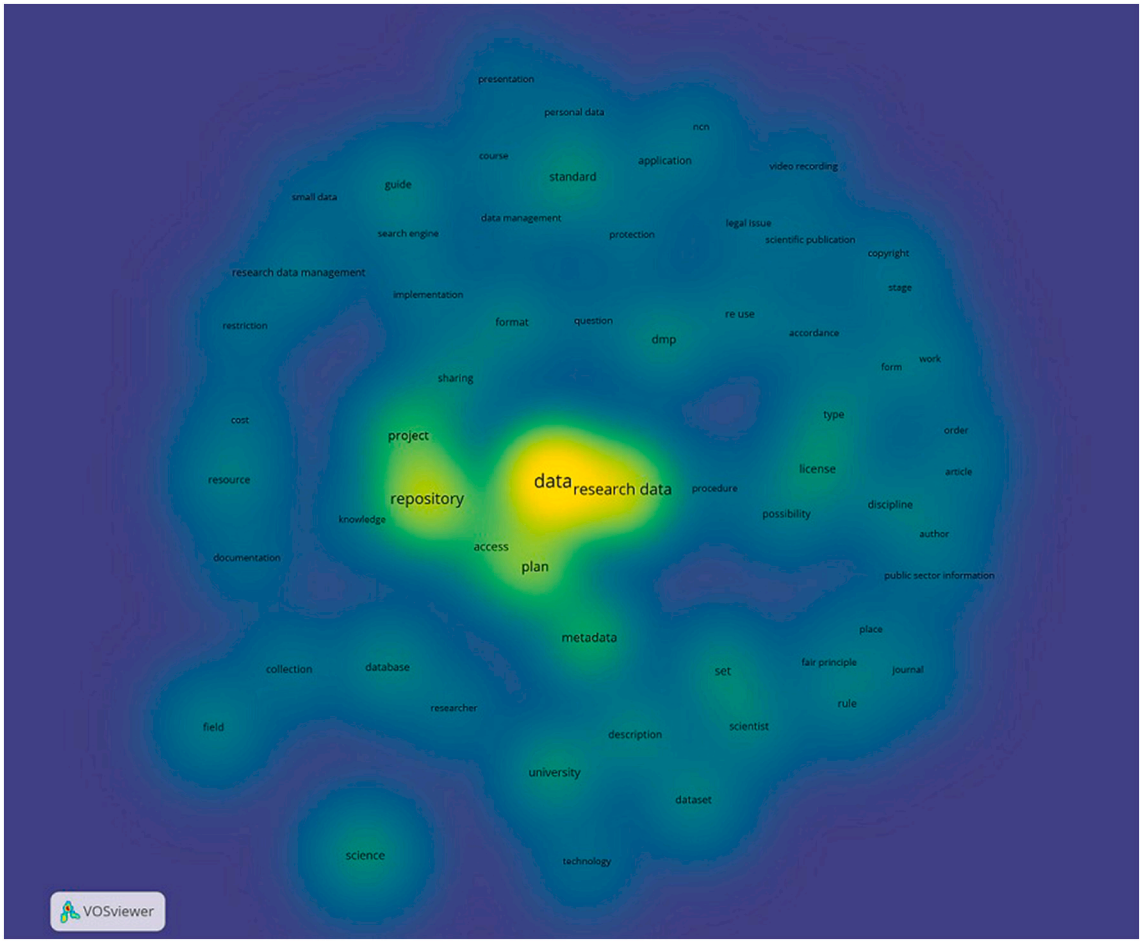

Fig. 6. Density visualization of websites content.

becomes. Conversely, the smaller the number of phrases in the vicinity of the point and the lower their weight, the closer to blue is the color. The map in Fig. 6 shows that the phrases "data", "research data", "repository", and "plan" have a higher density than others, which means that these phrases have strong relationships with others (Chen et al., 2016). The others have a similar density of occurrences.

Another way to present the most important issues raised on the library websites is to present their content in the form of a data cloud (Fig. 7) which was obtained with the help of the Voyant Tools program that provides multiple text mining functions (Gregory et al., 2022). This is a web tool that can analyze a single text document or several documents combined into one body. It allowed us to extract words repeated in the text with different frequencies and display them in different visualizations. For clarity of the image, a stop-word list was used that contains words that should be excluded from the results. Stop-word lists contain so-called function words that do not carry as much meaning, such as determiners and prepositions. The word cloud positions the words in the text so that the terms that occur the most frequently are positioned centrally and sized the largest. The color of words and their absolute position are not significant. Automatic text analysis allows for semi-automatic determination of its subject area, which in this case was the use of RDS in the libraries under study. Voyant Tools can also help validate the results of quantitative text analysis. The word cloud distribution indicated that words such as "research", "data", "management", "open", "metadata", "project", "use", "sharing", "access" appeared frequently, thus pointing to their importance.

Visualization is the greatest advantage of Voyant Tools, with over 25 different visualization formats to understand and explain the data contained in the text corpus. The most useful tools for the authors of this article were Cirrus, Summary, Trends, Reader, and Context.

The TermsBerry tool is also useful in the analysis, as it allows us to group the most frequent words in the form of a cluster, as well as to observe the co-occurrence relationship between them. This tool provides a way of exploring high-frequency terms and their collocates (words that occur in proximity). The highest-frequency words appear at the middle and in larger bubbles. The user can hover the cursor over the selected word, which turns green (Fig. 8) while all words in relation to it turn pink. Each bubbles indicate the collocate frequency for that word. The darker the pink, the more often the word is used. The number of occurrences common to the selected word is indicated below the related word. As shown in Fig. 8, the words "management", "repository/repositories" as well as "plan", "open", and "sharing" were in the closest relation to the keyword "data", the most frequent in the set.

The most important issues presented in the tag cloud confirmed the topics previously identified in the clustering process. After using a stop-word list, removing punctuation, and stemming, the five most common words in the corpus were "data" (1902), "research" (915), "repository/repositories" (381), "management" (293), and "open" (263). These are general terms but allowed us to define the subject area of the web resource. There were also less frequent, but important, terms such as "plan" (193), "scientific" (170), "metadata" (163), "sharing" (145), "access" (130), and "DMP" (91). The main issues were RDM and DMP as well as repositories and metadata for data present in the educational and training materials on the library websites.

Using the same tool, an attempt was made to estimate the maturity level of Polish academic libraries. For this purpose, the RDS maturity model by Cox et al. (2019) was used (Table 1). This model provides three clusters of expressions that characterize successive levels of RDS

Fig. 7. Tag cloud released with the Voyant Cirrus tool.

maturity. Compliance for cluster 1 means library responses to main drivers, such as funders and the continuation of existing practices. It is generally characterized by a formal policy coupled with advisory services. Stewardship for cluster 2 means continuation of traditional roles of libraries, but in relation to data repositories and journals and not to books and other publications. It is associated with the creation of a repository and associated services. Transformation (cluster 3) means deep transformation of library services to support high-level analytical activities. For each cluster, marked by a different color, the frequency of expressions (words or phrases) on the library web pages and the average of the occurrences were determined. Expressions related to the scope of the activities for each cluster, referenced in the model, were searched. The results obtained are visualized in Fig. 9.

The expressions in cluster two had the highest frequency on the library websites, whereas those in cluster three appeared with the lowest (minimum) frequency. Therefore, our task was to determine the expressions appearing on the tested websites of libraries. Their distribution at individual maturity level was determined by the adopted RDS maturity model (Cox et al., 2019). The level of maturity of the RDS surveyed (in this case, Polish academic libraries) is determined by the frequency of occurrence of a particular expression on library websites. Higher frequency of expressions in a given cluster provider that a specific level of maturity has been reached. In this research process, co-occurrence and other relations (visualized earlier) also were taken under consideration. In this way, the data obtained with the aid of web scraping and other NLP techniques were linked to the maturity model created by other methods.

Results and discussion

Libraries and other similar institutions offer very large volume of unstructured data (Zeng, 2019) and datafication means processing them not only into machine-readable materials (as in digitization), but also into machine-processable materials. They are the most varied in type, nature, and quality and also pose the most problems during processing. This study set out to analyze text documents available on library websites related to RDM. Under the conditions of textual processing of information, text matching was the main means of full text search for a long time. Recently, content-based search has been used in relation to text semantics. Instead of relying on full text searching, methods such as semantic-based analysis, extraction, mining, and tagging are used to facilitate the discovery of information from unstructured data.

The data collected from the websites of academic libraries using the web scraping technique and then analyzed with the aid of several NLP tools allowed for the determination of the RDS maturity level of these libraries and, thus, their ability to meet the information needs of their users in the field of research data services. The analysis of the website content of Polish academic libraries showed that the maturity of the RDS offered by these libraries was at good or average level, in line with the Cox et al. (2019) maturity model. The main issues raised on the websites concerned research data repositories, DMP creation, data storage, and metadata. On the other hand, there was almost no interest in new skills, placed in the third cluster of the model, such as visualization, integration, or data analysis, which, in opinion of Cox et al. (2019) demonstrate the high maturity level of RDS. These services were best developed in

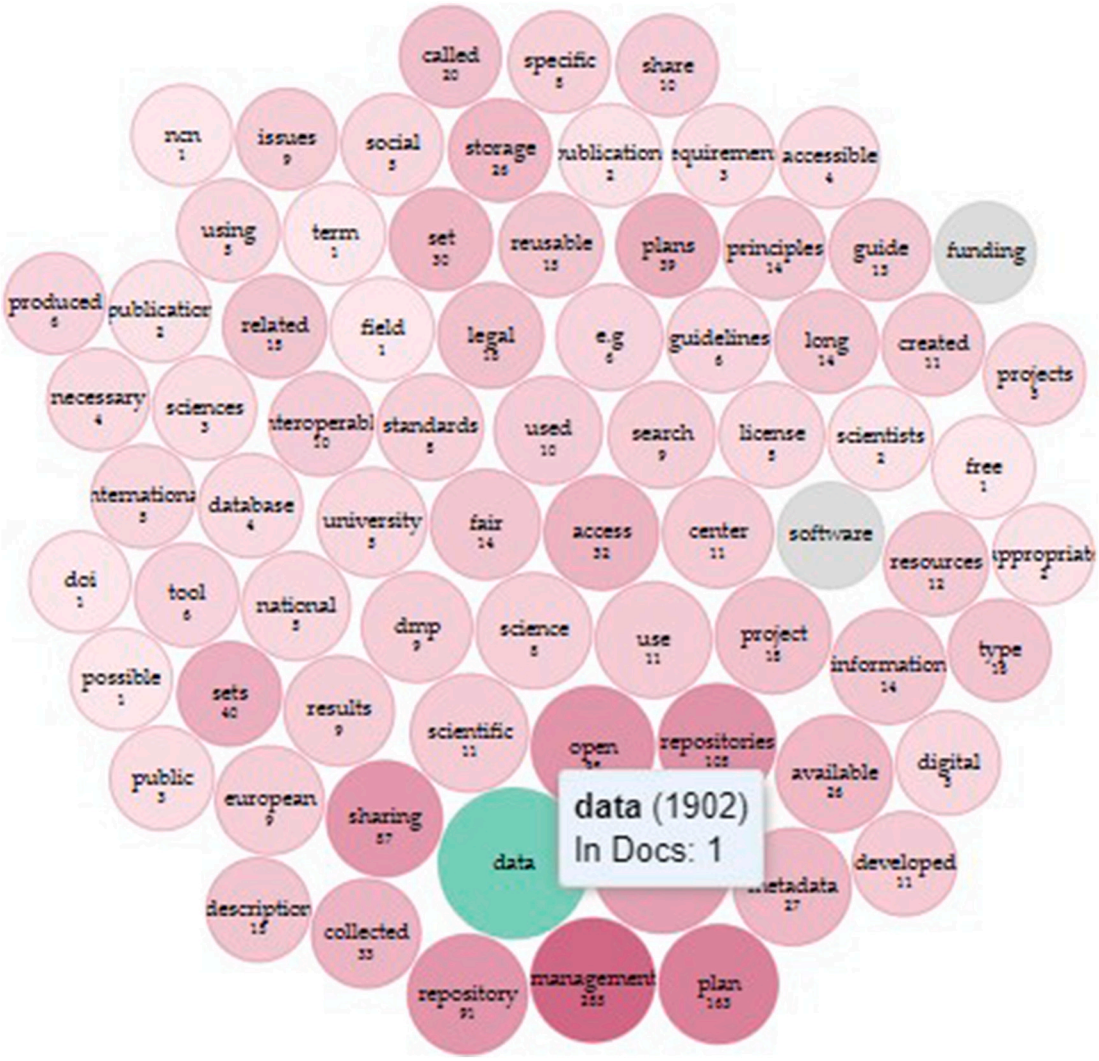

Fig. 8. Refined TermsBerry (Voyant tools).

large academic centers with the highest number of large universities. However, this is not a rule of thumb, as RDM initiatives are also being launched in smaller academic centers. It is also worth noting that, using web scraping, it was possible to select only about 14 % of libraries from the total number of these institutions taken into account. Other libraries do not provide any information on their websites regarding RDM services issues or at least web scraping was unable to locate them.

The level of RDM maturity is also related to the type of institution. Research has shown that RDM processes are most developed in university (classical) and technical university libraries. This confirms the hypothesis adopted about the correlation of the activity of libraries (visible on their web sites) with the activity of home universities in the field of research data, as RDM seems to play the most important role in research conducted at these types of universities. This is because the former include the largest universities with large libraries and extensive faculty, while the latter are the most technologically advanced. Numerous experimental studies are conducted on both. On the other hand, the lesser interest in RDM in libraries of medical and agriculture universities is disturbing, where the services on the library websites were often limited to placing a link to a national subject repository for medical sciences called the Polish Platform of Medical Research (PPM), which, incidentally, is an example of a well-functioning community cooperation.

Various types of data are used in the library RDS. Bibliographic data science uses metadata that should be regarded as structured data. In the study, we indicated the potential for analyzing the content of Web documents placed on library websites, which is closer to what is described as bibliomining, or data mining for libraries (Nicholson, 2003). Therefore, it can be concluded that, unlike structured data used in bibliographic data science, in this article we deal with unstructured data. Based on the big data methods described in the article, the integration process occurred, allowing the transition from unstructured data to structured data, whereby such data become 'smarter' (Simović, 2018; Zeng, 2019) and at the same time, data can be affected by the phenomenon of datafication. Big data technology has shown unique advantages in processing and analyzing unstructured data (Kitchen & McArdle, 2016). Information made available by the libraries may be regarded as 'small data', but following their aggregation they become big data. During these processes, the volume of big data is of lesser importance when compared to their value, as there is the possibility of arriving at important findings from such data on any scale, large or small (Schramm & Shafaghi, 2020). Such methodology differs significantly from the methods used so far in research on the maturity of RDS (e.g., Cox et al., 2017; Cox et al., 2019; Kim, 2021), which mainly used sociological methods, such as interviews, surveys, and/or observations. If web analysis was applied (e.g., Kouper et al., 2017; Singh et al., 2022), it was cast merely in an adjunct role.

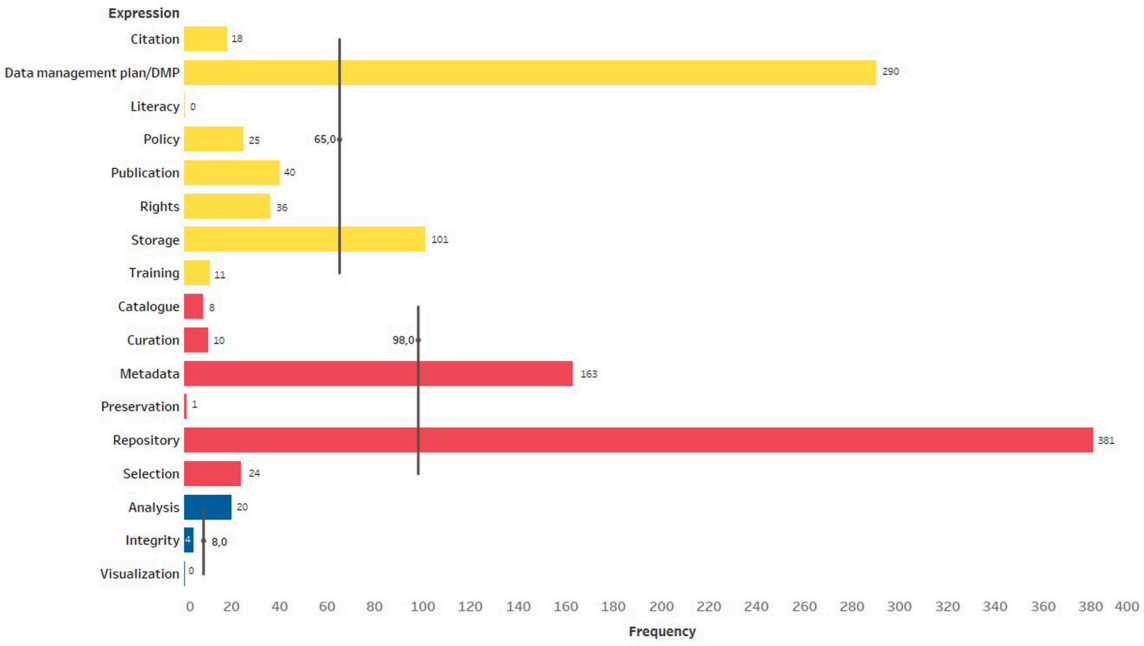

Fig. 9. Frequency of RDS expression clusters. Yellow: cluster 1 - Compliance: Translation of existing skills. Red: cluster 2 - Stewardship: Reskilling of existing staff. Blue: cluster 3 - Transformation: Acquisition of new skills. Vertical lines: average frequency in the cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Conclusions

The findings from this research indicate the value of using big data techniques in RDS maturity testing, such as web scraping, and analytical tools supporting NLP processes. This is inclusive of the tools available in the open-source mode. The continuation of research using these methods would allow determining the actual dynamics and geography of changes in the library RDS. Extending the research to other countries would allow for the benchmarking of the level of maturity of the service and, thus, also the level of its suitability to meet the needs of the information users. At the same time, overall usefulness of the Cox et al. (2019) maturity model in such studies was effectively corroborated.

The study findings indicated the lack of the most mature type of library activities, thereby pointing to the necessity of implementing significant changes in the organization of work. This merely goes to show that libraries tend to treat activities in the RDS area in a conservative manner by trying to have them adapted to their traditional functions. The libraries were inclined to create the types of services that corresponded to their existing activities (training, open access), and the already existing skill set of the librarians (repository, selection, access, curation, metadata). As a result, RDM was often interpreted through the "traditional" roles of libraries. In order to bring about truly meaningful change to this organizational paradigm, the librarians will be required to boost their analytical expertise relative to big data techniques as part of their current educational programs.

The study was limited by being confined to a single country and only one type of library (academic), so this factor should be taken into account in any subsequent studies by way of extending the geographic scope, and the type of libraries to be put under research scrutiny, as well as the actual dynamics of change over time. Other GLAM institutions may also be considered. However, despite these shortcomings, the present study points to the value of using big data tools and procedures in this type of analysis. Thanks to their application, a clear picture of RDS in Polish libraries was obtained, in particular, the level of their maturity and that of the management of research data in science. Understanding the geographic location and organizational diversity of RDS should help in learning more about the advantages and disadvantages of considering each one of these variables. It did assist in the identification of problems faced by the data librarians, while attempting to offer their users mature information services.

CRediT authorship contribution statement

Marek Nahotko: Conceptualization, Methodology, Writing – Orygi-nal draft, Writing – Review & editing, Funding acquisition.
Magdalena Zych: Software, Formal analysis, Writing.
Aneta Januszko-Szakiel: Writing, Validation.
Małgorzata Jaskowska: Investigation, Verification.

Funder

Jagiellonian University, grant no: ID.UJ DigiWorld 2021/1

Data availability

Data will be made available on request.

Research data set for publication "Big data-driven investigation of the maturity of library research data services (RDS)" (Original data) (Jagiellonian University Repository)

Appendix 1. Keyword list: Polish (original) and English (translated)

In color are the terms for which hits were obtained during web scraping

No	Polish keyword	English keyword
1	DMP	DMP
2	RDM	RDM
3	analiza danych badawczych	research data analysis
4	archiwizacja danych badawczych	research data preservation
5	baza danych badawczych	research data database
6	centrum zarządzania danymi	data management center
7	dane badawcze	research data
8	dane z badań	data from research
9	Data Management Plan	Data Management Plan
10	deponowanie danych badawczych	research data depositing
11	depozyt danych badawczych	research data deposit
12	długoterminowa archiwizacja danych badawczych	research data long-term archiving
13	długoterminowe składowanie danych badawczych	research data long-term storage
14	długotrwałe przechowywanie danych	research data long-term preservation
15	dostarczanie danych	data submission
16	dostęp do wyników badań	access to research results
17	dostępność danych badawczych	research data availability
18	ekosystem danych badawczych	research data ecosystem
19	FAIR Data	FAIR Data
20	generowanie danych	data generating
21	gospodarowanie danymi badawczymi	research data management
22	gospodarowanie zasobami danych badawczych	research data resource management
23	gromadzenie danych	data acquisition
24	identyfikator danych	data identifier
25	identyfikator dostępu do danych	data access identifier
26	interoperacyjność danych badawczych	research data interoperability
27	jakość danych badawczych	research data quality
28	karencja udostępniania danych badawczych	prolongation of research data sharing
29	kodowanie danych	data coding
30	kuratorstwo danych	data curation
31	metadane danych badawczych	research data metadata
32	nadzór nad danymi	data supervision
33	niszczenie danych	data destruction
34	ochrona użyteczności danych	data usability protection
35	odczyt danych	data reading
36	opiekun danych	data guardian
37	organizacja zasobów danych badawczych	organization of research data resources
38	otwarte dane badawcze	open research data
39	otwartość danych	research data openness
40	pielęgnowanie danych badawczych	maintenance of research data
41	plan zarządzania danymi	data management plan
42	polityka postępowania z danymi badawczymi	research data handling policy
43	polityka zarządzania danymi	data management policy
44	ponowne wykorzystanie danych badawczych	research data re-use
45	postępowanie z danymi	data handling
46	pozyskiwanie danych	data acquisition
47	program zarządzania danymi	data management program
48	przechowywanie danych	data storage
49	przetwarzanie danych badawczych	research data processing
50	publikowanie danych badawczych	research data publishing
51	Research Data Management	Research Data Management
52	rozpowszechnianie danych badawczych	research data dissemination
53	schemat metadanych danych badawczych	research data metadata schema
54	selekcja danych	data selection
55	składowanie danych	data storage
56	standard metadanych danych badawczych	research data metadata standard
57	strategia udostępniania danych badawczych	research data sharing strategy
58	strategia zarządzania danymi badawczymi	research data management strategy
59	system danych badawczych	research data system
60	system zarządzania danymi badawczymi	research data management system
61	trwała ochrona danych	permanent data protection
62	utrwalanie danych	data persistence
63	użyteczność danych	data usefulness
64	walidacja danych	data validation
65	wynik badań naukowych	research result
66	wytwarzanie danych badawczych	research data production
67	zabezpieczenie danych badawczych	research data securing
68	zarządzanie danymi z badań	research data management
69	zasób danych badawczych	research data resource
70	zamykanie danych	restricted data
71	zasady FAIR Data	FAIR principles
72	zbiory danych badawczych	research data collection

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.

Al-Jaradat, O. (2021). Research data management (RDM) in Jordanian public university libraries: Present status, challenges and future perspectives. *Journal of Academic Librarianship, 47*(5), Article 102378.

Amirian, P., van Loggerenberg, F., & Lang, T. (2017). Data science and analytics. In P. Amirian, T. Lang, & F. van Loggerenberg (Eds.), *Big data in healthcare* (pp. 15–37). Cham: Springer. Springer Briefs in Pharmaceutical Science & Drug Development.

Anderson, C. (2004). The long tail. *Wired Magazine, 12*(10). https://www.wired.com/2004/10/tail/.

Arbia, G., & Nardelli, V. *On Spatial Lag Models estimated using crowdsourcing, Web-scraping or other unconventionally collected data.* (2020). Retrieved from https://arxiv.org/abs/2010.05287 Accessed September 12, 2022.

Arzberger, P., et al. (2004). Promoting access to public research data for science, economic, and social development. *Data Science Journal, 3*, 135–152.

Ball, R. (2019). Big data and their impact on libraries. *American Journal of Information Science and Technology, 3*(1), 1–9.

Baškarada, S., & Koronios, A. (2017). Unicorn data scientist: The rarest of breeds. *Program, 51*(1), 65–74.

Berman, E. (2017). An exploratory sequential mixed methods approach to understanding researchers' data management practices at UVM: integrated findings to develop research data services. *Journal of eScience Librarianship, 6*(1), Article e1104.

Big data. (2022). Oxford English Dictionary. Retrieved from https://www.oed.com/view/Entry/18833#eid301162177 Accessed September 6, 2022.

Borgman, C., et al. (2016). Data management in the long tail: Science, software and service. *International Journal of Digital Curation, 11*(1), 128–149.

Brooks, D. (2013). The Philosophy of Data, The New York Times. Retrieved from: https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html Accessed September 8, 2022.

Bryant, R., Lavoie, B., & Malpas, C. (2017). *A tour of the research data management (RDM) service space. The realities of research data management.* Dublin, OH: OCLC Research.

Chen, X., et al. (2016). Mapping the research trends by co-word analysis based on keywords from funded project. *Procedia Computer Science, 91*, 547–555.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*(1), 21–26.

Cooke-Davies, T. (2004). Project management maturity models. In P. Morris, & J. Pinto (Eds.), *The Wiley guide to managing projects* (pp. 1234–1255). Hoboken: Wiley.

Cox, A., Kennan, M., Lyon, L., et al. (2019). Maturing research data services and the transformation of academic libraries. *Journal of Documentation, 75*(6), 1432–1462.

Cox, A., et al. (2017). Development in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology, 68*(9), 2182–2200.

Cox, A., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science, 46*(4), 299–316.

Crosby, P. (1979). *Quality is free*. New York: McGraw-Hill.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(5), 70–76.

Delserone, L. (2008). At the watershed: Preparing for research data management and stewardship at the University of Minnesota Libraries. *Library Trends, 57*(2), 202–210.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Diggle, P. J. (2015). Statistics: A data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 178*(4), 793–813.

Van Dijk, J. (2020). *The network society*. London: SAGE Publ.

Dongo, I., et al. (2020). Web scraping versus Twitter API: a comparison for a credibility analysis. In *Proceedings of the 22nd international conference on information integration and web-based applications & services* (pp. 263–273). https://doi.org/10.1145/3428757.3429104

Dumbill, E. (2013). Making sense of big data. *Big Data, 1*(1), 1–2.

Feger, S., et al. (2020). 'Yes, I comply!': motivations and practices around research data management and reuse across scientific fields. In *, 4. Proceedings of the ACM on human-computer interaction* (p. 141). New York: ACM. CSCW2.

Foreman, J. W. (2013). *Data smart: Using data science to transform information into insight.* Hoboken, NJ: John Wiley & Sons.

Gardner, S., Juricek, J., & Xu, G. (2008). An analysis of academic library web pages for faculty. *The Journal of Academic Librarianship, 34*(1), 16–24.

Garoufallou, E., & Gaitanou, P. (2021). Big data: Opportunities and challenges in libraries, a systematic literature review. *College & Research Libraries, 82*(3), 410–435.

Goczyła, K. (2021). Udanawianie wszystkiego. Retrieved from *Pismo PG, 28*(3), 79–80 Accessed October 12, 2022 https://pg.edu.pl/documents/1152961/104233222/202103.pdf.

Granville, V. (2014). *Developing analytic talent: Becoming a data scientist.* Hoboken, NJ: John Wiley and Sons.

Gregory, K., Geiger, L., & Salisbury, P. (2022). Voyant tools and descriptive metadata: A case study in how automation can compliment expertise knowledge. *Journal of Library Metadata, 22*(1–2), 1–16.

Gudivada, V., Rao, D., & Raghavan, V. (2015). Big data driven natural language processing research and applications. In V. Govindaraju, V. Raghavan, & C. Rao (Eds.), *33. Handbook of statistics* (pp. 203–238). Amsterdam, Oxford: Elsevier.

Harari, Y. N. (2017). *Homo deus: A brief history of tomorrow*. London: Vintage.

Hayashi, C. (1998). Preface. In C. Hayashi (Ed.), *Data science, classification, and related methods* (pp. V–VII). Tokyo: Springer Verl.. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996.

Heidorn, B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280–299.

Hoy, M. (2014). Big data: An introduction for librarians. *Medical Reference Services Quarterly, 33*(3), 320–326.

Humphrey, W. (1989). *Managing the software process*. Reading, MA: Addison-Wesley.

Iwasiński, Ł. (2020). Theoretical bases of critical data studies. *ZIN Information Studies, 58*(1A), 96–109.

Karasti, H., Baker, K., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work, 15*(4), 321–358.

Kelleher, J. D., & Tierney, B. (2018). *Data science*. Cambridge, MA.: MIT Press.

Kim, J. (2021). Determining research data services maturity: The role of library leadership and stakeholder involvement. *Library and Information Science Research, 43*(2), Article 101092.

Kim, S., & Syn, S. (2021). Practical considerations for a library's research data management services: The case of the National Institutes of Health library. *Journal of the Medical Library Association, 109*(3), 450–458.

Kitchen, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society, 3*(1), 1–10.

Koltay, T. (2017). Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science, 49*(1), 3–14.

Kouper, I., Fear, K., Ishida, M., et al. (2017). Research data services maturity in academic libraries. In L. Johnston (Ed.), *1. Curating research data* (pp. 153–170). Chicago: ACRL. Practical strategies for your digital repository.

Lahti, L., Marjanen, J., Roivanen, H., & Tolonen, M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly, 57*(1), 5–23.

Laney, D. (2001). 3-D data management: controlling data volume, velocity and variety. META Group Research Note. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf Accessed September 22, 2022.

Laskowski, C. (2021). Structuring better services for unstructured data: Academic libraries are key to an ethical research data future with big data. *Journal of Academic Librarianship, 47*(4), Article 102335.

Leetaru, K. (2015). Mining libraries: Lessons learned from 20 years of massive computing on the world's information. *Information Services & Use, 35*(1–2), 31–50.

Lewis, M. (2010). Libraries and the management of research data. In S. McKnight (Ed.), *Envisioning future academic library services* (pp. 145–168). London: Facet Publ.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics, 11*(5/6), 295–324.

Li, J., Lu, M., Dou, G., & Wang, S. (2017). Big data application framework and its feasibility analysis in library. *Information Discovery and Delivery, 45*(4), 161–168.

Liu, S., & Shen, X. (2018). Library management and innovation in the big data era. *Library Hi Tech, 36*(3), 374–377.

Lunn, S., Zhu, J., & Ross, M. (2020). Utilizing web scraping and natural language processing to better inform pedagogical practice. In *2020 IEEE Frontiers in Education Conference (FIE), Uppsala, 21-24 Oct. 2020* (pp. 1–9). IEEE. https://doi.org/10.1109/FIE44824.2020.9274270 Accessed August, 15, 2022.

Manjunatha, K. (2016). Content analysis of special library websites: An analytical study. *International Journal of Next Generation Library and Technologies, 2*(2), 1–9.

Manyika, J., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. San Francisco, CA: McKinsey Global Institute.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data. The essential guide to work, life and learning in the age of insight.* London: John Murray.

McCaffrey, M., & Giesbrecht, W. (2016). Teaching data librarianship to LIS students. In L. Kellam, & K. Thompson (Eds.), *Databrarianship: The academic data librarian in theory and practice* (pp. 355–373). Chicago: ACRL.

Naur, P. (1974). *Concise survey of computer methods*. Lund, Sweden: Petrocelli Books.

Nicholson, S. (2003). Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research works. *Journal of the American Society for Information Science and Technology, 54*(12), 1081–1090.

Osika, G. (2021). Dilemmas of social life algorithmization – Technological proof of equity. Scientific papers of Silesian University of Technology. *Organization and Management Series, 151*, 525–538.

Pareek, S., & Gupta, D. (2012). Information about services and information resources on websites of selected libraries in Rajasthan: A study. *DESCIDOC Journal of Library & Information Technology, 32*(6), 499–508.

Petrovich, E. (2021). Science mapping and science maps. *Knowledge Organization, 48*(7/8), 535–562.

Pinfield, S., Cox, A., & Smith, J. (2014). Research data management and libraries: relationships, activities, drivers and influences. *PLoS ONE, 9*(12), Article e114734.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51–59.

Qin, J., Crowston, K., Flynn, C., et al. (2015). *A capability maturity model for research data management.* Syracuse, NY: School of Information Studies, Syracuse University.

Qin, J., Crowston, K., Flynn, C., & Kirkland, A. (2017). Pursuing best performance in research data management by using the capability maturity model and rubrics. *Journal of eScience Librarianship, 6*(2), Article e1113.

Ran, C., Yang, L., & Hu, L. (2021). Revisit the implementation status of research data management in Chinese academia. *Journal of Academic Librarianship, 47*(3), Article 102350.

Ratner, B. (2017). *Statistical and machine-learning data mining: Techniques for better predictive modelling and analysis of big data*. Boca Raton: Chapman and Hall/CRC Press.

Regueira, U., Alonso-Ferreiro, A., & Da-Vila, S. (2020). Women on YouTube: Representation and participation through the web scraping technique. *Comunicar, 28*(63), 31–40.
Reinhalter, L., & Wittman, R. (2014). The library: Big data's boomtown. *Serials Librarian, 67*(4), 363–372.
Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big data and data science in critical care. *Chest, 154*(5), 1239–1248.
Schramm, M., & Shafaghi, M. (2020). Moving from big data to smart data for enhanced performance, business efficiency, and new business models. *Journal of International Business and Management, 3*(2), 1–17.
Schutt, R., & O'Neil, C. (2014). *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly Media.
Semeler, A., Pinto, A., & Rozados, H. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science, 51*(3), 771–780.
Simović, A. (2018). A big data smart library recommender system for an educational institution. *Library Hi Tech, 36*(3), 498–523.
Singh, R., Bharti, S., & Madalli, D. (2022). Evaluation of research data management (RDM) services in academic libraries of India: A triangulation approach. *Journal of Academic Librarianship, 48*(6), Article 102586.
Song, I. Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems, 33*(4), 364–373.
Stanton, J. M., et al. (2012). Interdisciplinary data science education. In *Special issues in data management* (pp. 97–113). Washington, DC: American Chemical Society (ACS Symposium Series, 1110).
Tang, R., & Hu, Z. (2019). Providing research data management (RDM) services in libraries: Preparedness, roles, challenges, and training for RDM practice. *Data and Information Management, 3*(2), 84–101.
Tella, A., & Kadri, K. (2021). Big data and academic libraries: Is it big for something or big for nothing? *Library Hi Tech News, 38*(2), 15–23.
Tenopir, C., Sandusky, R., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research, 36*(2), 84–90.
Tenopir, C., et al. (2017). Research data services in European academic research libraries. *LIBER Quarterly, 27*(1), 23–44.
Tenopir, C., et al. (2015). Research data services in academic libraries: Data intensive roles for the future? *Journal of eScience Librarianship, 4*(2), Article e1085.
Tiwari, A., & Madalli, D. (2021). Maturity models in LIS study and practice. *Library and Information Science Research, 43*, Article 101069.
Tolonen, M., Marjanen, J., Roivainen, H., & Lahti, L. (2019). Scaling up bibliographic data science. In C. Navaretta, M. Agirrezabal, & B. Maegaard (Eds.), *Digital humanities in the Nordic countries* (pp. 450–456). Copenhagen: Univ. of Copenhagen. Proc. of the digital humanities in nordic countries 4th conference.
Tu, Y., Chang, S., & Hwang, G. (2021). Analysing reader bahaviours in self-service library stations using a bibliomining approach. *The Electronic Library, 39*(1), 1–16.
Uhlir, P., & Schröder, P. (2007). Open data for global science. *Data Science Journal, 6*(special issue), 36–53.
Uzun, E. (2020). A novel web scraping approach using the additional information obtained from web pages. *IEEE Access, 8*, 61726–61740.
Verbaan, E., & Cox, A. (2014). Occupational sub-cultures, jurisdictional struggle and third space: Theorising professional service responses to research data management. *Journal of Academic Librarianship, 40*(3–4), 211–219.
Virkus, S., & Garoufallou, E. (2019). Data science from a perspective of computer science. In E. Garoufallou, F. Fallucchi, & W. De Luca (Eds.), *Metadata and semantic research. 13th international conference, MTSR 2019*. Cham: Springer Verl.
Voulgaris, Z. (2014). *Data scientist: The definitive guide to becoming a data scientist*. Westfield, NJ: Technics Publications.
Wainer, H. (2015). *Truth or truthiness: Distinguishing fact from fiction by learning to think like a data scientist*. Cambridge, MA.: Cambridge University Press.
Whyte, A., & Tedds, J. (2011). *Making the case for research data management. DCC briefing papers*. Edinburgh: Digital Curation Centre.
Whyte, A. (2014). A pathway to sustainable research data services: From scoping to sustainability. In G. Pryor, S. Jones, & A. Whyte (Eds.), *Delivering research data management services* (pp. 59–88). London: Facet.
Xiu, J., & Wang, M. (2014). Competencies and responsibilities of social science data librarians: An analysis of job descriptions. *College & Research Libraries, 75*(3), 362–388.
Xu, Z., et al. (2022). A scoping review: Synthesizing evidence on data management instruction on academic libraries. *Journal of Academic Librarianship, 48*(3), Article 102508.
Yan, Y. (2020). Industry requirements for translators across China before COVID-19: Analyzing 51job listings through web scraping. *Revista Argentina de Clínica Psicológica, 29*(4), 768–779.
Yang, Z., Zhu, R., & Zhang, L. (2016). Research on the capability maturity model of digital library knowledge. In B. Xu (Ed.), *Proceedings of the 2nd International Technology and Mechatronics Engineering Conference (ITOEC 2016)* (pp. 333–337). Zhengzhou: Atlantis Press. Chongqing, China, May 21–22, 2016.
Yidavalapati, J., Sinha, P., & A, S. (2021). Research data management and services in South Asian academic libraries. *Library Philosophy and Practice*, 6457.
Yoon, A. (2017). Role of communication in data reuse. *Proceedings of the Association for Information Science and Technology, 54*(1), 463–471.
Yoon, A., & Donaldson, D. (2019). Library capacity for data curation services: A US national survey. *Library Hi Tech, 37*(1), 811–828.
Yoon, A., & Schultz, T. (2017). Research data management services in academic libraries in the US: A content analysis of libraries' websites. *College & Research Libraries, 78*(7), 920–933.
Zeng, M. (2019). Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. *El Profesional de la Información, 28*(1), Article e280103.
Zhan, M., & Widén, G. (2019). Understanding big data in librarianship. *Journal of Librarianship and Information Science, 51*(2), 561–576.