Research data repositories in the RDM cycle: challenges and strengths for curators/data stewards

by Juan-José Boté-Vericad, Ana Carballo-Garcia,
Mònica Bautista-Villaescusa, Sharon Healy

Introduction

The management and publication of research results are crucial components of contemporary academic research. Academic institutions must ensure effective organization, preservation, and accessibility of research findings. In this context, data repositories have gained significant importance by providing specific digital infrastructure to store, preserve, and facilitate access to research[1]. Some repositories are discipline-specific, targeting researchers from specific academic fields, while others are institutional, serving the research community affiliated with a particular university or other academic institution[2].

These repositories exist to meet the growing demand for efficient and sustainable data management solutions that ensure data integrity, accessibility, and compliance with data sharing requirements. They promote transparency, collaboration, and long-term preservation of valuable research outcomes[3]. Therefore, it is essential to address key issues and to understand how the curators of these repositories fit into the research data landscape.

JUAN-JOSÉ BOTÉ-VERICAD, Universitat de Barcelona, Facultat d'Informació i mitjans audiovisuals & Centre de recerca en Informació, comunicació i cultura, e-mail: juanjo.botev@ub.edu.
ANA CARBALLO-GARCIA, Universitat de Barcelona, e-mail: anacarballogarcia@gmail.com.
MÒNICA BAUTISTA-VILLAESCUSA, Universitat de Barcelona, e-mail: mbautivi36@alumnes.ub.edu.
SHARON HEALY, Maynooth University, Department of Computer science, e-mail: schealy.ire@gmail.com.

1 Juan-José Boté; Miquel Termens, *Reusing data technical and ethical challenges*, «DESIDOC Journal of library and information technology», 39 (2019), n. 6, p. 329-337, DOI: 10.14429/djlit.39.06.14807.

2 Juanjo Boté; Julià Minguillón, *Preservation of learning objects in digital repositories*, «Revista de Universidad y Sociedad del Conocimiento», 9 (2012), n. 1, p. 217-230, DOI: 10.7238/rusc.v9i1.1036.

3 Susanne Blumesberger, *Repositorien als Tools für ein umfassendes Forschungsdatenmanagement: am Beispiel von PHAIDRA an der Universitätsbibliothek Wien*, «Bibliothek Forschung und Praxis», 44 (2020), n. 3, p. 503-511, DOI: 10.1515/bfp-2020-2026.

In November 2022, the European Commission's Directorate-General for research and innovation published the *European research data landscape*[4], the final report on a study conducted between 2021 and 2022 by various organizations and companies throughout Europe specializing in research data management. The aim of the research was to characterize the research ecosystem in Europe, focusing on researchers and research data repositories. A survey was sent to over 840.000 researchers, and responses were received from 15.066 individuals, while 316 research data repositories were also surveyed. The study analyzed data production, consumption, and deposition practices, as well as researchers' knowledge and application of FAIR (findable, accessible, interoperable, and reusable) principles. The results showed that a relatively low percentage of researchers deposit their data in repositories. The report highlighted the importance of research institutions in supporting data management and sharing, as researchers rely on institutional support services. The need for clear policies, training, and guidance in data management and FAIR principles was emphasized, and recommendations were provided to improve research data management, including local support, comprehensive assistance throughout the data lifecycle, FAIR assessment, and raising awareness of the benefits of FAIR implementation. The role of information professionals, particularly data stewards, was considered crucial for implementing these recommendations and effectively managing data repositories within research institutions.

Curators play a fundamental role in managing and preserving research data in academic environments. Their responsibilities include data organization, metadata creation, data preservation, and ensuring data quality and accessibility[5]. To successfully manage the repositories in their care, they need specialized data management skills and knowledge in library science[6].

In this study, we focus on curators and their experiences to shed light on their responsibilities, competencies, training needs, and perceptions. This will contribute to the advancement of research data management practices in academic institutions. We also consider open access policies and the integration of FAIR principles[7], which are important elements of research data management[8]. Additionally, we examine the use of shared data in repositories and the challenges associated with their daily operation and development.

The dynamic nature of research data management requires continuous exploration and investigation. Understanding the work of data curators in repositories is crucial to optimize data management practices, tackle challenges, and seize opportunities in this constantly evolving field.

4 European Commission. Directorate-General for research and innovation, *European research data landscape: final report*. Luxembourg: Publications Office of the European Union, 2022, DOI: 10.2777/3648.

5 Annette Strauch, *To begin, at the beginning [...]: Bibliotheken als „Player" im professionellen Forschungsdatenmanagement*, «Bibliothek Forschung und Praxis», 44 (2020), n. 2, p. 166-169, DOI: 10.1515/bfp-2020-0025.

6 Glyneva Bradley-Ridout, *Preferred but not required: examining research data management roles in health science librarian positions*, «Journal of the Canadian Health Libraries Association», 39 (2018), n. 3, p. 138-145, DOI: 10.29173/JCHLA29368.

7 Ana Carballo; Juan-José Boté-Vericad, *Fair data: history and present context*, «Central European journal of educational research», 4 (2022), n. 2, p. 45-53, DOI: 10.37441/cejer/2022/4/2/11379.

8 Juan-José Boté-Vericad; Sharon Healy, *Academic libraries and research data management*, «Vjesnik bibliotekara Hrvatske», 65 (2022), n. 3, p. 171-193, DOI: 10.30754/vbh.65.3.1016.

Purpose of the study

The purpose of this study is to examine the roles of curators in academic research data repositories, and their perceptions regarding skills and training, research data management (RDM) practices, and the limitations, strengths, and opportunities they encounter in relation to their work. In addition, the study examines the functions of academic research repositories across five global regions, and the extent to which FAIR principles are integrated into repository strategies.

Specifically, we seek to answer the following research questions:
- RQ1: what training or skills do data curators need to better perform their tasks?
- RQ2: how do data curators assist researchers with RDM?
- RQ3: how is dataset reuse managed in repositories?

This paper is structured in five sections. Following this introductory section, section 2 presents a review of the literature on RDM and research data repositories. Section 3 then describes the methods and materials used in this study, and section 4 presents the results and provides an analysis of the main findings. Finally, section 5 offers a discussion of the results and some conclusions.

Literature review

Research data management

The effective management of research data is essential for academic institutions to support the research lifecycle and enable the dissemination of findings. To this end, academic libraries have increasingly taken on the role of providing RDM services to support researchers and their data needs. This section examines various studies conducted on RDM services provided by academic libraries in different regions, including Australia, Ireland, the UK, the US, and Canada. These studies identify common challenges faced by academic libraries in providing RDM services, such as understanding researchers' data storage needs, the complexity of RDM, and the costs associated with it. However, they also highlight opportunities for academic libraries to enhance their support throughout the research lifecycle and to promote their value to campus communities.

Concerning services offered by academic libraries, Corrall, Kennan and Afzal[9] performed a study examining planned services, target audiences, service constraints, and staff training needs, with a sample of 140 participating academic libraries. The findings showed Australia had the highest percentage (77,1%) for bibliometrics training, while Ireland had the highest percentages for citation reports (88,9%) and research impact calculation (77,8%). The UK had the highest percentage in RDM services, including technology infrastructure (53,8%), institutional repositories or access to external datasets (37,5%), and RDM guidance (41,3%). Australia was the leader in RDM guidance (25,7%) and planning (21,2%), albeit with lower rates compared to other library services.

Tenopir [*et al.*][10] investigated RDM practices and policies in US and Canadian academic research libraries. Using a stratified random sample of 351 library directors (response rate: 63%), the study found that 49,5% of libraries offered dataset location services, while only 23,5% provided consultation on data standards. Additionally,

9 Sheila Corrall; Mary Anne Kennan; Waseemal Afzal, *Bibliometrics and research data management services: emerging trends in library support for research*, «Library trends», 61 (2013), n. 3, p. 636-674, DOI: 10.1353/lib.2013.0005.

10 Carol Tenopir [*et al.*], *Research data management services in academic research libraries and perceptions of librarians*, «Library and information science research», 36 (2014), n. 2, p. 84-90, DOI: 10.1016/j.lisr.2013.11.003.

26,3% assisted with faculty data management plans, and 53,6% collaborated with other RDS providers. Training opportunities included conferences (62%), courses (53%), and library-based training (32%). The study concluded that libraries were increasingly supporting research data management and planning.

In relation to policies, Cox and Pinfield[11] examined UK libraries' involvement in RDM through a questionnaire sent to research and higher education institutions. According to the 81 responses, RDM open access, policy, and copyright advice were the most widely implemented services. Challenges included understanding researchers' data storage needs, library workforce skills, and RDM costs. Cox, Pinfield and Smith[12] conducted 26 interviews with librarians responsible for RDM, addressing 16 'wicked' problems. Key findings of their study highlighted RDM's complexity and the need for information professionals to develop skills and attitudes to tackle the challenges it poses in their training. Another study[13] explored RDM experiences of 36 academic library professionals in the United States. Through qualitative data collection via interviews and focus groups, they identified various factors influencing RDM, such as technical resources and the researchers' perceptions about the library. Their findings revealed that 31% of academic library professionals had no RDM experience, while 46% of them had only some. Digital repositories and dedicated RDM experts offered important support for researchers. However, librarians' expertise and services were not perceived as highly qualified by researchers. The study highlighted opportunities for librarians to enhance their support throughout the research lifecycle and promote the library's value to the campus community.

Bishop [*et al.*][14] conducted an interview study with librarians (N=10) and research integrity officers (N=12) in the United States to explore their roles in research data services. Gaps in RDM were identified across institutions. Researchers managed their own RDM needs after initial training, while librarians could collaborate and support faculty, students, and staff. Research services from other units were not integrated into research centers. Both groups expressed the need for improved cyberinfrastructure, increased budgets and staffing, and additional training to enhance RDM support. Yoon and Schultz[15] analyzed 185 academic library websites in the US to examine their promotion of RDM services. Their results showed that libraries primarily offered data management (65%) and data services (17%), with limited emphasis on data curation (3%). Most libraries provided data deposit services (60%) and data man-

11 Andrew M. Cox; Stephen Pinfield, *Research data management and libraries: current activities and future priorities*, «Journal of librarianship and information science», 46 (2014), n. 4, p. 299-316, DOI: 10.1177/0961000613492542.

12 Andrew M. Cox; Stephen Pinfield; Jennifer Smith, *Moving a brick building: UK libraries coping with research data management as a 'wicked' problem*, «Journal of librarianship and information science», 48 (2016), n. 1, p. 3-17, DOI: 10.1177/0961000614533717.

13 Ixchel Faniel; Lynn Silipigni Connaway, *Librarians' perspectives on the factors influencing research data management programs*, «College and research libraries», 79 (2018), n. 1, p. 100-119, DOI: 10.5860/crl.79.1.100.

14 Wade Bishop [*et al.*], *Potential roles for science librarians in research data management: a gap analysis*, «Issues in science and technology librarianship», 98 (2021), DOI: 10.29173/ISTL2602.

15 Ayoung Yoon; Teresa Schultz, *Research data management services in academic libraries in the US: a content analysis of libraries' websites*, «College and research libraries», 78 (2017), n. 7, p. 920-933, DOI: 10.5860/crl.78.7.920.

agement planning (80%), but educational programs were less common (34,5% offered workshops/lectures). Only 21,8% provided training on data sharing and reuse in their data management training. The study recommended increased engagement of academic libraries with the development of educational services for RDM.

Research data management initiatives
Concerning RDM initiatives, Bardyn [*et al.*][16] share their experience with the translational research and information lab (TRAIL) at the University of Washington (UW) Health Services Library in the United States, addressing faculty members' ongoing needs for research data management and preservation services. The aim of the initiative was to coordinate data and innovation services for clinical researchers by partnering with various units at UW, guided by principles of collaboration, quality, assessment, diversity, education, and access. This partnership expanded the university's clinical research data management (CRDM) services, including data visualization, survey creation, bioinformatics consultation, and the use of emerging technologies.

Read [*et al.*][17] present a model for RDM implementation in six US academic libraries. Motivations included enhancing library visibility, understanding institutional research, and addressing data management needs. The participating libraries were at University at Buffalo, University of Delaware, Drexel University, Duquesne University, Stony Brook University, and Temple University. An eight-module pilot training program was developed. In these modules, participants were required to use two out of three components: 1) a template and strategies for data interviews; 2) a teaching tool kit to teach an introductory RDM class; and/or 3) strategies for hosting a data class series. The study highlighted the importance of building communities to support librarians in updating their RDM skills.

More recently, Kim and Syn[18] have examined the case of the National Institutes of Health (NIH) Library in the United States. In four interviews with librarians who provide data services, they utilized a two-model crosstab framework combining three categories from the Online Computer Library Center (OCLC) RDM service categories and six levels of the data lifecycle. The OCLC categories were education services, expert services, and curation services, while the data lifecycle categories encompassed data creation, description, storage, sharing, and preservation. This framework helped identify service gaps, and the results revealed collaboration at multiple levels, involving various NIH units and external partners.

Concerning RDM literacy, Steiner[19] examined Lincoln University in New Zealand and its information literacy program on RDM. Through interviews with six head

16 Tania P. Bardyn [*et al.*], *Health sciences libraries advancing collaborative clinical research data management in universities*, «Journal of eScience librarianship», 7 (2018), n. 2, p. 1-14, DOI: 10.7191/jeslib.2018.1130.

17 Kevon B. Read [*et al.*], *A model for initiating research data management services at academic libraries*, «Journal of the Medical Library Association», 107 (2019), n. 3, p. 432-441, DOI: 10.5195/jmla.2019.545.

18 Soojung Kim; Sue Yon Syn, *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 (2021), n. 3, p. 450-458, DOI: 10.5195/jmla.2021.995.

19 Katrin Steiner, *Forschungsdatenmanagement und Informationskompetenz – Neue Entwicklungen an Hochschulbibliotheken Neuseelands: Research data management and information literacy – New developments at New Zealand university libraries*, «Information – Wissenschaft & Praxis», 66 (2015), n. 4, p. 230-236, DOI: 10.1515/iwp-2015-0040.

research officers from academic libraries, barriers to RDM were identified, including the diversity of researchers' interests and challenges in data sharing. Collaboration among academic libraries was hindered by university low funding and publish-or-perish attitude. The study compared the situation with the one in Germany, highlighting the importance of a funding policy by the German Research Foundation (DFG) to promote open access and facilitate the implementation of local RDM initiatives.

Finally, Lindstädt and Schmit[20] explain how a consortium was founded by ZB MED and 20 German institutions to create a national research data infrastructure for life sciences. The consortium aimed at ensuring data quality and interoperability through standardized practices. They introduced PUBLISSO, a support service for open access publishing in fields like medicine, biology, and agriculture. Libraries can provide comprehensive support, including publication services and research data management. For instance, Humboldt-Universität-Berlin created a guide for using DMPonline, a tool from the Digital Curation Center, to develop data management plans in the German-speaking world.

Data stewards and data curation
Tammaro [*et al.*][21] investigate the evolving socio-technical practice of data curation, which involves technical systems and services structured around the research data life cycle, as well as a range of social activities aimed at community building. The research employed a mixed method approach with a combination of quantitative and qualitative strategies. The study participants were recruited from 24 organizations in nine countries, and the data were gathered through interviews and content analysis of job announcements. The study highlights common themes in social aspects of research data management, particularly promoting open data awareness, fostering a data-sharing culture, and supporting researchers in data-intensive environments. Despite more than a decade of data curation research and practice, there is still no real consensus on the terminology and titles for people who are involved in providing research data management services. The study concludes that data curation is emerging as a hybrid profession that combines technical and public services skills, and the core responsibilities including outreach, training, and advocacy reflect the influence of librarianship on the role of the data curator.

Also, according to Orrù[22], data stewards are responsible for managing and maintaining data over time. They work with data creators, users, and curators to ensure that data are properly documented, stored, and made accessible. Data stewards also develop policies and procedures for data management, monitor compliance with those policies, and provide training and support to researchers on data management best practices. Overall, data stewards play an important role in ensuring the long-term usability and preservation of research data.

20 Birte Lindstädt; Jasmin Schmit, *Das Management von Forschungsdaten als Handlungsfeld wissenschaftlicher Bibliotheken: forschungsunterstützung am Beispiel ZB MED – Informationszentrum Lebenswissenschaften*, «Bibliothek Forschung und Praxis», 43 (2019), n. 1, p. 42-48, DOI: 10.1515/bfp-2019-2006.

21 Anna Maria Tammaro [*et al.*], *Data curator's roles and responsibilities: an international perspective*, «Libri», 69 (2019), n. 2, p. 89-104, DOI: 10.1515/libri-2018-0090.

22 Damiano Orrù, *Open data steward: bibliotecari e alfabetizzazione ai dati aperti*, «AIB studi», 60 (2020), n. 2, p. 311-323, DOI: 10.2426/aibstudi-12123.

According to Gruber [*et al.*][23], data stewardship is a crucial aspect of managing research data, and there are several competencies that are essential for effective data stewardship. These competencies include data management, data curation, data documentation, data quality control, data security and privacy, and data sharing and dissemination. Data stewards should have a strong understanding of these competencies and be able to apply them throughout the data lifecycle. Data management involves collecting, storing, processing, and analyzing data, and data stewards should be familiar with relevant tools and technologies for these activities. Data curation involves organizing and preserving data for long-term use, and data stewards should be familiar with the best practices for managing and sharing data. Data documentation involves creating documentation that accurately describes data, and data stewards should ensure that documentation provides the context, content, and structure of the data. Data quality control involves ensuring that data are accurate, complete, and reliable, and data stewards should be able to assess and improve data quality. Data security and privacy involve protecting data according to applicable laws and regulations, and data stewards should be familiar with relevant principles and best practices. Data sharing and dissemination involve sharing data to maximize their impact and reuse while respecting data owners' rights and interests, and data stewards should identify appropriate channels for sharing data.

Dataverse: the standardized software for research data
Dataverse is becoming a standardized solution for research data repositories, particularly in academic and research institutions. Dataverse provides an open-source, web-based platform for researchers to share, preserve, and cite their research data. It offers features such as version control, data citation, and persistent identifiers, which are important for ensuring the long-term accessibility and usability of research data. Dataverse also provides integration with other tools and services, making it easy to deposit and share research data across multiple platforms. As an increasing number of institutions and publishers adopt data sharing policies, Dataverse is likely to keep growing in popularity as a standardized solution for research data repositories.

Park[24] emphasizes the importance of data sharing in research and provides guidance on how to deposit data into the Harvard Dataverse repository. The *World Journal of Men's Health* has adopted a clinical data sharing policy and uses Harvard Dataverse as its repository. Park suggests that scientific journal editors should choose an appropriate platform and take part in the practice of data sharing. The author concludes that data sharing is essential for research and predicts that data sharing will become a widespread trend for scientific and medical journals across all fields.

Chen, Chiu and Cline[25] explore the Dataverse global research data management consortium and the university libraries that take part in it. They surveyed 13 top US

---

23 Alexander Gruber [*et al.*], *Kompetenzen von Data Stewards an österreichischen Universitäten*, «Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare», 74 (2021), n. 1, p. 12-32, DOI: 10.31263/voebm.v74i1.6255.

24 Hyun Jun Park, *How to share data through Harvard Dataverse, a repository site: a case of the World Journal of Men's Health*, «Science editing», 9 (2022), n. 1, p. 85-90, DOI: 10.6087/kcse.270.

25 Hsin-liang Chen; Tzu-Heng Chiu; Ellen Cline, *Academic libraries and research data management: a case study of Dataverse global adoption*, «Information discovery and delivery», 51 (2023), n. 2, p. 166-178, DOI: 10.1108/IDD-04-2022-0028.

research universities and examined common data management practices among institutions participating in Dataverse. They also conducted a literature review on scholarly communication and research data. The majority of Dataverse members are research universities or institutions, with English being the primary language (used by around 70% of members). Dataset and collection growth has been mostly minimal or flat, with only a few members seeing significant use of their data portals. The authors recommend further research into data discovery and metadata implementation, as well as library research data services and research data management policy. They suggest that research data management policy and research data services mutually support each other, and that understanding how institutions and researchers work with data will be important for helping future library users in those areas. They conclude that the lessons learned from the Dataverse project could assist other research data initiatives and academic library services. The Dataverse case study provides insights into research data management and could be used to improve research data initiatives and academic library services.

Methodology

This section outlines the approach and methodology used in this study for sample selection, participant recruitment, questionnaire design, and data analysis.

Sample selection and recruitment

Curators of academic research repositories constituted the target audience for this study. To select the sample of repositories, the first source was the Re3data website <https://www.re3data.org/>, which serves as a comprehensive directory of research data repositories. The main purpose of the Re3data.org directory is to provide a reliable resource for researchers and other data users. The criteria include aspects such as data policies, access options, metadata standards, and long-term data preservation.

The next step involved choosing the repositories is based on the following conditions: a) the repository must be managed by one institution and not under a consortium; b) the repository must specifically cover one of five global regions (Africa, the Americas, Asia, Europe, or Oceania); c) the repository must make the metadata for their datasets visible; and d) the repository must provide a contact address.

After establishing a convenience sample of 19 repositories that met the criteria outlined above, we emailed the contact addresses. In the email we described the purpose of the study, provided a copy of the interview questions, and kindly asked email recipients to participate in the study in one of the following ways: - typing up their answers in a text document and sending the file to us; - sending us a voice message in an MP3 file; - replying to the email with their answers to our questions; or - filling out the answers in an online form.

For the online form, we used EU Survey, an open-source tool for conducting surveys and the «official survey management tool»[26] of the European Commission.

Questionnaire

The questionnaire consisted of 14 questions, divided into three parts. In part 1, participants were asked about their gender, education, skills and training, the specific

---

26 European Commission, *About EUSurvey*. 2023, <https://ec.europa.eu/eusurvey/home/about>.

type of training or skills they would need to perform their roles more effectively, and the obstacles they encountered in the development of the repository. In part 2, participants were questioned about their institution and repository, specifically in terms of the type of training and guidelines they offer to assist researchers with their RDM practices and the use of the repository. Participants were also asked whether they collaborate with other organizations or contract external services for assistance in the RDM data lifecycle. In addition, they were asked about the research data they manage and the strategies implemented (if any) to ensure research integrity and the integration of FAIR principles. Part 3 focused on how participants reuse the research data from their repository, including questions about whether they use any tools to track usage and citation of the research data, whether they collect usage, download, and upload statistics to measure user trends and patterns, and whether they compile any information on which fields or disciplines have the most data uploaded to the repository.

More specifically, the questionnaire contained the following questions:
- What type of background, skills and training do you have?
- What types of skills and training would benefit your performance in this role?
- What types of obstacles do you encounter in the development of the repository and its day-to-day operations?
- Do the researchers at your institution receive training in any/all of the following topics?
- What type of training (seminars/face-to-face/tutorials/videos)?
- Does your repository provide guidelines or training to researchers on how to prepare and upload their research data to the repository? If so, what type of guidelines/training do you provide (e.g., seminars/face-to-face meetings/tutorials/videos)?
- Does your repository partner with any other organizations, or hire the services of any other organizations to assist with the RDM lifecycle of the data (e.g., management, storage, preservation, provision of access, etc.), including organizations that provide technical software or infrastructure? If so, please provide some examples.
- What type of research data is deposited in the repository? Does your repository guarantee that the research data deposited complies with the FAIR principles? If so, what measures are taken to ensure this?
- Do you compile any information on how the research data in your repository is reused for other research studies or investigations? If so, please provide examples.
- Does the repository use any tools for tracking the use and citation of the data provided?
- Are usage, download and upload statistics collected on the repository, and are user usage trends and patterns tracked?

Data analysis
In total, 19 respondents participated in the study. We obtained answers by various means: in a text file as a response to the email (N=2), using the EU Survey form (N= 17), and recorded interviews (N=2). Two responses were removed from the final analysis because they were incomplete. The recorded interviews were transcribed verbatim using Whisper open-source software[27]. We also received email responses indicating that the institution could not participate in the study for various reasons (N=10). These responses were entered in a field diary and integrated into the results in a manner

---

similar to that used by Villarroya and Boté-Vericad[28]. All answers were compiled in text form for further study, using inductive analysis[29] which involves allowing the codes to emerge from the participants' answers. The data was then analyzed by two coders. To validate the code reliability, both researchers codified a sample individually and an intercoder reliability test was performed using Krippendorff's Alpha methodology[30]. The result was α=0,835, indicating a high degree of consistency between observers.

The participating repositories are shown below in Figure 1.

Universidad Nacional de Rosario (Argentina)	Laboratory for research of individual differences at the University of Belgrade, Faculty of Philosophy (Serbia)
KU Leuven (Belgium)	Nanyang Technological University (Singapore)
Pacific Salmon Foundation (Canada)	DataFirst, University of Cape Town (South Africa)
Universidad del Rosario (Colombia)	University of Pretoria (South Africa)
Technische Universität Darmstadt (Germany)	Aston University (United Kingdom)
University of Stuttgart (Germany)	Johns Hopkins University (United States of America)
Ludwig-Maximilians-Universität München (Germany)	University of Arizona (United States of America)
Central University of Haryana (India)	Purdue University (United States of America)
Università degli studi di Milano Statale (Italy)	International Research Institute for climate and society, Climate School, Columbia University (United States of America)
Università degli studi di Padova (Italy)	

Figure 1 – Participant repositories

Results and analysis
The participants' responses contained a diverse range of answers. Although there were some similar responses, others reflected considerable disparity, leading to inconclusive results. The gender distribution of the participants was relatively even, with eight females, ten males, and one participant choosing not to disclose their gender. Notably, there was no significant pattern suggestive of a gender gap in data repository management.

However, the data obtained on the backgrounds and competencies of the participants reveal a wide range of backgrounds and experiences. The largest proportion of participants had a scientific background in STEM (N=9) or other disciplines related to science, such as psychology, computer science, engineering, physics, sociology, or library science. Others had extensive experience in research data management, metadata preservation,

28 Anna Villarroya; Juan-José Boté-Vericad, *The gender and LGBTQ perspectives in library and information science: a case study at the University of Barcelona*, «Library and information science research», 45 (2023), n. 2, p. 1-9, DOI: 10.1016/j.lisr.2023.101238.

29 Philipp Mayring, *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim; Basel: Beltz, 2010.

30 Klaus Krippendorff, *Content analysis: an introduction to its methodology*. London: Sage, 2004.

digital preservation, geospatial analysis, and software development. Three participants indicated they had a PhD, while others had training or experience in specific tools, languages, or principles, such as open science, Python programming, or FAIR principles. This highlights the importance of having a multidisciplinary team with a variety of skills and backgrounds when working on research projects. It should be noted that the number of information professionals with a background in library science was much lower (N=4) than those from technical disciplines. However, respondents with LIS training indicated extensive professional experience in the sector as well as complementary training in skills related to programming or data management.

The participants identified a range of competencies and skills as necessary for the effective performance of the role of data repository manager. These have been divided into four main categories based on the possibility of establishing relationships between them:

1. Technical/Advanced user skills:
- Emphasis on programming, software engineering, and computer engineering skills.
- Proficiency in the use of specialized software such as GeoTools.
- Ability to handle less popular operating systems such as Linux in all its distributions.
- Deep knowledge of SQL databases for advanced data management.
2. Data analysis & management skills:
- Skills related to ingesting, processing, interpreting, and visualizing system data.
- Knowledge of technologies related to the semantic web (linked open data).
3. Information literacy & librarianship skills:
- Knowledge of intellectual property.
- Proficiency in information retrieval techniques.
- Familiarity with information literacy and bibliographic citations.
4. Transversal skills:
- Competencies applicable to other fields of a multidisciplinary nature.
- Organizational skills, expertise, aptitude for collaborative learning, and capacity for lifelong learning, given the constantly changing nature of RDM.

This information can be useful for determining training needs and developing professional development plans for individuals or teams. It also highlights the importance of continuous learning and development in the rapidly evolving field of research data management. Figure 2 provides a summary of the areas in which respondents identified the need for skills. In addition to the areas shown in Figure 2, 'research process' skills are also in demand among STEM professionals.

Technical/ Advanced user	Data analysis & management	Information literacy & librarianship	Transversal
Computer science	Dataset connection	Information and library science	Expert advice
Coding	Data analysis tools	Librarian training in Spanish	Peer learning
GeoTools	Data curation	Information retrieval	Organizational
Linux	Data science	Scientific processes	Ongoing learning
SQL databases	Data visualization	Usage and citations	
Technical	Linked open data	Intellectual property laws	
	RDM training		
	Semantic metadata		
	Standards		

Figure 2 – Areas in which respondents identified the need for skills

Participants highlighted various challenges faced in the development and daily operation of repositories. The issue identified most often was a lack of time and resources, which had a significant impact on the work individuals were able to do. This constraint was often correlated with a shortage of administrative or support personnel for data processing. Some respondents emphasized the challenge posed by a lack of technical knowledge or support, indicating a need for specialists in data processing, particularly data scientists, at institutions maintaining repositories. Several specific challenges were also identified, including difficulties in preserving datasets for various subjects, in keeping databases and hyperlinks up to date, and in navigating legal and ethical issues. Other issues mentioned included the reluctance of some researchers to share their data and the lack of incentives or awareness to encourage researchers to deposit their data into repositories.

The integration of the repository from a technical perspective was also highlighted as a major challenge. This issue was linked to the need for skills in computer science and software development. Additionally, participants mentioned a lack of awareness among researchers of the importance of entering complete metadata in the data entry process. Overall, the obstacles described underscore the need for sufficient time, resources, technical expertise and awareness, in order to overcome challenges in repository development and operation. Hiring specialized personnel and promoting the value of the repository are strategies suggested to address these obstacles effectively.

On the other hand, several responses highlighted researchers' lack of knowledge regarding how to deposit items in the institutional repository, as well as their misconceptions about its purpose. This often leads to misunderstandings or situations where researchers choose not to share their data or knowledge openly due to either a lack of understanding or insufficient incentives (N=5). Alongside the need for increased resources and technical knowledge for managing repositories, respondents emphasized the need to promote a better understanding of the research cycle and intellectual property as fundamental aspects of responsible management.

To address these challenges, it is crucial to provide researchers with clear guidance and education about the repository's purpose and benefits. Efforts should be made to improve researchers' awareness and understanding of the repository's role in fostering open sharing and collaboration. It is also essential to allocate adequate resources for repository management, including technical expertise and support. Educating researchers about the research process and intellectual property rights could contribute to fostering a culture of conscious data management and effective utilization of institutional repositories.

Curators of research data repositories face a myriad of challenges impeding development and growth. The lack of guidance, the absence of incentives, and insufficient technical skills are some of them. This Singaporean female curator and her Canadian male counterpart highlight the global significance of finding solutions to these pressing issues:

1. lack of guidance for research data repositories development.
2. It requires dedicated time for development.
3. No incentives available for development and growth.
4. Various ethical concerns.
5. Lack of research data knowledge in the research and academic domain.
6. Requirement of research data literacy. (Curator, female, Singapore).

Insufficient technical skills in a rapidly-changing data management environment[...] No one to consult with, requiring self-reliant study using Internet[...]

> Maintaining currency of database content and hyperlinks[...] Delays in creation
> of data-sharing agreements sometimes involving lawyers, any violations of
> which would be difficult to enforce. (Curator, male, Canada).

Researchers at the institutions surveyed receive training on various aspects of research
data management, including open science principles, metadata standards, FAIR
principles, and data management planning. The training programs are offered in a
variety of formats, such as seminars, in-person training sessions, webinars, workshops,
video tutorials, and self-paced online modules. Some institutions specifically cater
their training to postgraduate research students, while others offer tailored programs
for different faculties. However, it is important to note that most of these training
programs are optional, and participation in face-to-face sessions tends to be lower
compared to online sessions.

Of the different types of training offered at institutions, webinars emerged as the
most common format. The training focused primarily on RDM, open access, data
management plans (DMPs), and FAIR data.

To further enhance research data management practices, institutions could con-
sider implementing more comprehensive, mandatory training programs for
researchers. Encouraging greater participation in face-to-face training sessions and
promoting the benefits of in-person interactions would also be beneficial. Additionally,
exploring innovative training approaches, such as gamification or peer-to-peer learn-
ing, could foster greater engagement and knowledge retention among researchers.

Curators from various countries share insights on the challenges and training
initiatives related to research data management.

> Non-mandatory central courses on research data management including FAIR
> principles and metadata standards and data management planning. Currently,
> those courses are remote seminars. (Curator, male, Germany).

> There are many groups on campus offering different kinds of training so it's
> difficult to know what researchers actually receive. From the repository per-
> spective, our trainings include all of the topics mentioned in the question text.
> The training consists of workshops and web resources for self-learning. (Curator,
> male, United States).

> I train data entry persons in use of a metadata editor (jNAP) in creating iso-19115
> metadata. (Curator, male, Canada).

> Open Science principles for research data: seminars, face-to-face, data manage-
> ment planning: seminars, face-to-face. (Curator, female, Italy).

Most repositories also offer guidelines or training to researchers on how to prepare
and upload their research data. These resources come in various forms, including
online forms, workshops, web-based tutorials, written guides, seminars, in-person
training sessions, and video tutorials. Some repositories go beyond guidelines and
offer free RDM training courses specifically designed for postgraduates, or research
data literacy courses for researchers and students. Others provide manuals or extensive
online support to assist users in navigating the repository effectively. Certain repos-
itories even recommend the use of specific tools, such as the DataCite metadata gen-
erator, to ensure data quality and integrity during the uploading process. Overall,

the design of the repositories reflects an understanding of the importance of aiding researchers to manage and upload their data correctly. However, despite the availability of such training resources, the development and day-to-day operation of repositories continues to be negatively impacted by the limited time, technical knowledge, and experience that their users have. Addressing this challenge may involve providing researchers with more accessible and streamlined training options, leveraging user-friendly tools and interfaces, and fostering a supportive environment that encourages researchers to engage with repository training and utilize its resources effectively. We highlight a best practice from a curator from Argentina.

> Documentation is available on the RDA-UNR information website: <https://data-verse-info.unr.edu.ar/>. Curatorial consultation hours are available through virtual meetings with a specialized curator. (Curator, female, Argentina).

Responses to the question of whether repositories partner with other organizations to assist with the RDM data lifecycle revealed a range of different approaches. While some repositories handle the entire RDM lifecycle internally, others engage in partnerships with external organizations. These partnerships serve various purposes, such as the provision of managing infrastructure and technical software. Some respondents mentioned collaborating with their university's IT department to meet the repository's infrastructure needs. Others formed partnerships with external entities such as the World Bank, Dataverse Community, and the Leibniz Supercomputing Center. Such collaborations can provide expertise and resources in specific areas of RDM.

Several respondents also mentioned securing funding from programs such as EOSC Future and the Research data alliance (RDA). These funds are typically used to enhance the repository's adherence to key principles such as user orientation, sustainability, searchability, and interoperability. In short, the responses indicated widespread recognition of the value of partnerships to support different aspects of research data management. These partnerships enable repositories to effectively manage, store, preserve, and provide access to research data, leveraging the expertise and resources available through external organizations.

Repositories receive a diverse range of research data, encompassing various formats and types. Accepted data types include spreadsheets, interviews, databases, computer code, peer-reviewed and non-peer-reviewed data, institutional reports, audio, video, and images. Data formats supported include CSV, XLSX, ODT, PDF, source code, Docker containers, simulation runs, and hdf5 files, among others. There are generally no strict restrictions on the types of data that can be published, except for certain repositories that may exclude manuscripts, reports, or standalone theses. The primary consideration is to avoid including personal information, although large datasets may also pose problems of suitability for certain repositories. By providing flexibility in accommodating a wide range of data types and formats, repositories aim to encourage researchers to share and preserve their research outputs effectively. A curator from Belgium stated:

> The repository is open to all types of research data used and produced by researchers. There are no strict limits on what types of data are allowed to be published. The only limitation being [that] it should be research data or code. (Curator, female, Belgium).

The survey responses indicate different approaches among repositories to FAIR principles. 13 respondents explicitly stated that they strive to ensure that all research

data deposited complies with FAIR principles. Conversely, eight repositories do not offer any specific guarantee in this regard. Repositories that aim for FAIRness employ various measures to achieve compliance. These measures include data preservation to ensure long-term accessibility, requiring mandatory metadata to enhance findability, providing persistent and unique identifiers to facilitate proper citation and attribution, ensuring machine-readable metadata for interoperability, supplying licenses for data reuse, and adhering to international standards.

In addition to these measures, some repositories may require researchers to convert data to specific file formats, provide detailed information about the data creation process, and disclose any dependencies associated with the data. However, it is important to note that not all repositories have implemented comprehensive measures to ensure FAIRness. The degree of adherence to FAIR principles may vary across repositories, and some may still be in the process of developing and implementing strategies to enhance FAIRness in their data management practices.

In relation to adhering to the FAIR principles, three curators stated different positions, considering that they all make efforts to do so:

> We promote FAIR principles, and ensure that some mandatory fields have to be filled when depositing a dataset. This includes, a title, data collection period, keywords, and contact info in case of data requests/queries. All entries are reviewed and any changes which make the deposit more FAIR are encouraged. Unfortunately not all who deposit are keen to add any more time to provide more/better descriptive metadata regarding the dataset and at times we have to accept datasets we would like to have seen improved further as we think ultimately something is better than nothing. Further work is admittedly needed to encourage greater descriptive data regarding dataset[s] which will aid utility for the user. (Curator, male, United Kingdom).

> We assign persistent identifiers at both the dataset and individual data file level. We create comprehensive metadata at the project and data level, including DOIs and ORCIDs, and automatically identify the data format. The metadata can be harvested in different formats such as JSON-LD, Dublin Core, and RDFa. We are registered with DataCite, re3data, and fairsharing. (Curator, female, Argentina).

> We strive to comply with FAIR principles, but we do not provide guarantee. We curate every single dataset that is being deposited to the repository, and liaise with depositor to ensure that the dataset is FAIR. (Curator, male, Singapore).

Some repositories actively track the use of research data by means of various metrics, including citations, downloads, and related publications. Some have also reported the receipt of anecdotal evidence of citations and new collaborations resulting from the reuse of their datasets. Notably, one repository highlighted that resources deposited in their repository are frequently reused by researchers at all career stages and students from universities within their country or region. However, it is important to acknowledge that not all repositories track data reuse, and some rely solely on download counts as a measure. Tracking data reuse can be challenging as researchers often do not provide direct links to their data-based publications, making it difficult to establish a clear connection between data downloads and subsequent reuse.

Furthermore, tracking data reuse can be a time-consuming task, and it may be challenging to determine how data are reused once they are downloaded. Despite

these limitations, repositories continue to explore methods to enhance tracking mechanisms and gain deeper insights into the reuse of research data. In short, while there are repositories that actively monitor and observe instances of data reuse, acquiring a comprehensive understanding of the extent and impact of data reuse remains a complex and evolving process.

Regarding the use of data collected in repositories, responses varied in terms of the types of usage statistics tracked. Some respondents indicated that they collect information on downloads, uploads, and overall usage trends, while others do not track usage data. Certain repositories have their own integrated analytics systems, while others rely on external tools such as Google Analytics and DataCite for tracking usage. Based on the responses, download statistics appear to be the main metrics currently tracked. However, some respondents expressed plans to expand their tracking capabilities to include more comprehensive usage statistics in the future. It is worth noting that limited resources may prevent some repositories from conducting in-depth analysis of data reuse.

Although instances of research data reuse are observed in repositories, the extent of data reuse is not widely studied and depends on the specific repository and the nature of the data it holds. It is important for repositories to continue exploring ways to improve data tracking and analysis to gain valuable insights into data reuse patterns and to promote the use of research data resources.

Discussion and conclusions

This paper explores the challenges faced by research data repositories and data curators in the RDM cycle. In response to RQ1, this study has identified a pressing need to upgrade the digital skills of librarians. A large proportion of repository managers have backgrounds in different STEM disciplines, which gives them certain advantages over professional librarians, especially in terms of technical skills that enable them to understand how the products they manage work and how they can be improved with technological solutions (APIs, for example) to better meet users' needs.

This raises questions about the effectiveness of the training provided to librarians by official organizations and associations. Various analyses of the websites of organizations such as IFLA[31] (International Federation of Library Associations and Institutions), SEDIC[32] (Spanish Society for Scientific Documentation and Information), COBDC (the Official Association of Librarians-Documentalists of Catalonia), and COBDCV (the Official Association of Librarians-Documentalists of Valence) suggests that they are not providing in-depth training in a number of the skills identified in Figure 1, especially advanced training in programming or API integration.

However, the training section of the SEDIC website contains information on seminars in which tools such as DSpace, semantic web, and web scrapers are explained. The training offered by the same organization on data stewardship and advanced data management is also worth noting, although it is more explanatory than practical.

Data analysis and data management skills are of course also necessary. It should be noted that RDM involves not only data ingestion, but also data interpretation, visualization, and communication. This study has thus also identified the need to disseminate the FAIR data philosophy to professionals outside library and information

---

31 See <https://www.ifla.org/events/ifla-it-open-webinar-series-library-research-data-management-services-where-are-we-now/>.

32 See <https://www.sedic.es/data-stewardship-gestion-datos-investigacion/>.

science, who are less likely to be aware of its importance. These professionals also need to develop skills related to information literacy and intellectual property, in which librarians are much better trained. These findings are similar to those of Strauch[33], who points out that RDM is not consistent across organizations, especially in the case of health science librarians, and argues that RDM should be present in education programs. Read [*et al.*][34] found that responsibility for research data services is unclear within institutions and is predominantly placed on individual researchers or research teams. A number of the skills identified as necessary in the present study are not being addressed by institutions because of this lack of clarity.

In response to RQ2, there are a variety of RDM activities with which data curators assist researchers. Such activities include face-to-face training in the form of workshops or courses, which cover the entire RDM lifecycle to enable researchers to understand the full picture. Willems[35] describes an initiative to train librarians in the RDM lifecycle because research data services need to prioritize the RDM agenda[36].

These activities demonstrate data curators' understanding of the importance of helping researchers to manage and correctly upload their data. However, despite the availability of training resources, there are significant obstacles to the development and day-to-day operation of repositories, such as lack of time, technical knowledge, and experience of data curators with RDM.

Additionally, as reflected in the responses obtained, curators and researchers have to deal with a wide range of data types. This requires a high degree of flexibility to accommodate a wide range of data formats, as repositories aim to encourage researchers to share and preserve their research outputs effectively. This wide range of formats may explain why data curators' have low level of engagement in data deposit[37].

Repository users also need to be aware of the importance of inputting data and metadata according to minimum quality standards. The survey responses make it clear that there is a lack of awareness of the relevance of these standards among repository users, who demand incentives to perform these tasks.

The repositories surveyed have differing approaches to FAIR principles. Some repositories explicitly strive to ensure that the research data deposited comply with FAIR principles, while others do not provide any specific guarantee in this regard.

Greater awareness and understanding of the FAIR data philosophy, along with corresponding financial and governmental support, could overcome some of the difficulties currently faced by repositories, which could also lead to a change in the perception institutional users have of this powerful tool.

In response to RQ3, there is a notable lack of reuse and tracking of data in the repositories surveyed in this article. The return on investment of these digital spaces cannot be determined because there is no comprehensive tracking of institutional scientific output. Such tracking would increase the institutional and social benefits of these repositories and demonstrate the importance of open access in research data

33 A. Strauch, *To begin, at the beginning [...]* cit.

34 K. B. Read [*et al.*], *A model for initiating research data management services at academic libraries* cit.

35 Linda Willems, *Librarians share their top tips for research data management*, «Elsevier connect», April 20TH, 2023, <https://www.elsevier.com/connect/library-connect/librarians-share-their-top-tips-for-research-data-management>.

36 See note 15.

37 See note 13.

policies. Similar conclusions were drawn by Khan, Thelwall and Kousha[38], who found that repository managers of Dataverse repositories struggled to track dataset citations. However, solutions seem to be emerging, such as Provenance aware synthesis tracking architecture (PASTA) to track citations in documents from datasets[39].

On the other hand, to complement this work and reduce the intensity of the workflow that data curators and managers currently have to deal with, science communicators could constitute an excellent support. Communicating science makes it possible to give visibility to scientific production and scientific events taking place at academic institutions in a simpler and more enjoyable way. Institutional activities can be communicated on blogs or via social media. A good practical example is the case of the Argentinian repository that has a section in its blog called *Historias de datos*[40] (Data stories), where selected datasets are explained to the public in an interview format. The visibility of the research conducted is extremely important for researchers. The involvement of science communicators could encourage producers to publish their work with quality metadata, a practice that would facilitate subsequent processing by managers.

In conclusion, research data repositories are still in need of technological solutions to track the aggregation and reuse of datasets. Repositories could also benefit from support from scientific communicators in order to increase the visibility of their dataset. At the educational level, the professional profile of librarians may need to be updated, particularly to include more technical knowledge. However, it is important to avoid overburdening a single individual with all the responsibilities for RDM. Instead, a collaborative approach is needed, with recognition of the contributions of librarians to ensure the effectiveness of their work.

Finally, it is important to highlight that academic institutions cannot assume the whole burden of adhering to FAIR and open access data policies on their own. Governments and higher institutions should also support this task by providing resources to ensure that institutions can implement these policies without impediments such as those described in this article (lack of time, staff, technological resources, etc.). At the same time, in a globalized paradigm where science knows no physical boundaries, it is imperative that measures be taken to bridge gaps between countries, with special attention to developing countries.

38 Nushrat Khan; Mike Thelwall; Kayvan Kousha, *Are data repositories fettered? A survey of current practices, challenges and future technologies*, «Online information review», 46 (2022), n. 3, p. 483-502, DOI: 10.1108/OIR-04-2021-0204.

39 Mark Servilla [*et al.*], *The contribution and reuse of LTER data in the Provenance aware synthesis tracking architecture (PASTA) data repository*, «Ecological informatics», 36 (2016), p. 247-258, DOI: 10.1016/j.ecoinf.2016.07.003.

40 See <https://dataverse-info.unr.edu.ar/>.

ABSTRACT

JUAN-JOSÉ BOTÉ-VERICAD, Universitat de Barcelona, Facultat d'Informació i mitjans audiovisuals &
Centre de recerca en Informació, comunicació i cultura, e-mail: juanjo.botev@ub.edu.
ANA CARBALLO-GARCIA, Universitat de Barcelona, e-mail: anacarballogarcia@gmail.com.
MÒNICA BAUTISTA-VILLAESCUSA, Universitat de Barcelona, e-mail: mbautivi36@alumnes.ub.edu.
SHARON HEALY, Maynooth University, Department of Computer science, e-mail: schealy.ire@gmail.com.

Gli archivi dei dati della ricerca nel ciclo RDM: sfide e punti di forza per curatori/data steward
Questo articolo si occupa degli archivi accademici di dati della ricerca, e dei loro curatori, in diverse
località sparse su cinque regioni globali. Con l'ausilio di un campione costruito su misura, esploriamo
i ruoli dei curatori negli archivi di dati della ricerca e le loro impressioni su competenze e formazione,
su pratiche e limiti nella gestione dei dati della ricerca (RDM), sui punti di forza e le opportunità con le
quali si confrontano nel loro lavoro. Contestualmente lo studio esamina le funzioni svolte dagli archivi
accademici della ricerca e indaga fino a che punto i principi FAIR vengano effettivamente integrati nelle
strategie di archiviazione. I nostri risultati mostrano che, mentre certi archivi vengono gestiti da consorzi
di istituzioni per l'educazione superiore e centri di ricerca, in alcuni paesi essi sono ancora agli inizi,
spesso hanno una rappresentatività solo parziale, con una singola istituzione alla guida. Per di più, i
curatori si scontrano con il fatto che i ricercatori corredano i set di dati depositati con informazioni poco
chiare sul loro riuso. Nonostante gli archivi di dati siano costruiti per aderire ai principi FAIR, siamo
convinti che ci sia ancora molto lavoro da fare per insegnare ai ricercatori come usarli.

Research data repositories in the RDM cycle: challenges and strengths for curators/data stewards
This paper is concerned with academic research data repositories, and their curators, in different
locations across five global regions. Using an ad-hoc sample, we explore the roles of curators in
research data repositories and their perceptions of skills and training, research data management
(RDM) practices, and the limitations, strengths, and opportunities they encounter in relation to their
work. In addition, the study examines the functions of academic research repositories and the extent
to which FAIR principles are integrated into repository strategies. Our results show that while some
repositories are operated by consortia of higher education institutions and research centers, in some
countries research data repositories are still in their infancy, often with only a limited representation
led by one institution. Moreover, curators encounter problems with how researchers submit datasets
with unclear information for the reuse of the data. We argue that although data repositories are built
to comply with FAIR principles, there is still a lot of work to be done in relation to training researchers
on how to use data repositories.