Published in Online Information Review, DOI: 10.1108/OIR-04-2021-0204

Are data repositories fettered? A survey of current practices, challenges, and future technologies

Nushrat Khan[1], Mike Thelwall[2] and Kayvan Kousha[3]

[1]*n.j.khan@wlv.ac.uk*, [2]*m.thelwall@wlv.ac.uk*, [3]*k.kousha@wlv.ac.uk*
[1, 2, 3]University of Wolverhampton, Wulfruna St, Wolverhampton, WV1 1LY (United Kingdom)
[1]UCL Great Ormond Street Institute of Child Health, 30 Guilford St, London, WC1N 1EH (United Kingdom)

Abstract

Purpose
The purpose of this study is to explore current practices, challenges, and technological needs of different data repositories.

Design
An online survey was designed for data repository managers and contact information from the re3data data repository registry was collected to disseminate the survey.

Findings
In total 189 responses were received, including 47% discipline specific and 34% institutional data repositories. 71% of the repositories reporting their software used bespoke technical frameworks, with DSpace, EPrint, and Dataverse being commonly used by institutional repositories. 32% of repository managers reported tracking secondary data reuse while 50% would like to. Among data reuse metrics, citation counts were considered extremely important by the majority, followed by links to the data from other websites and download counts. Despite their perceived usefulness, repository managers struggle to track dataset citations. Most repository managers support dataset and metadata quality checks via librarians, subject specialists or information professionals. A lack of engagement from users and a lack of human resources are the top two challenges, and outreach is the most common motivator mentioned by repositories across all groups. Ensuring FAIR data (49%), providing user support for research (36%) and developing best practices (29%) are the top three priorities for repository managers. The main recommendations for future repository systems are - integration and interoperability between data and systems (30%), better research data management tools (19%), tools that allow computation without downloading datasets (16%) and automated systems (16%).

Originality
This study identifies the current challenges and needs for improving data repository functionalities and user experiences.

Keywords

Data repository, data reuse, impact measure, re3data.

Introduction

Sharing research data helps reproducible science by allowing research to be checked and letting other researchers exploit the data for new purposes. Data repositories play an important role for this by providing the policy and infrastructure to support effective curation and long-term preservation. Nevertheless, the development and implementation of repositories has been sporadic and varies between disciplines, with genomics, chemical crystallography, and biodiversity

extensively sharing open data (Faniel and Yakel, 2017; Robinson-García *et al.*, 2016; Khan *et al.*, 2021). Hence the disciplinary data repositories in these areas seem to be relatively mature and robust in terms of technology and policy. For example, the Global Biodiversity Information Facility (https://www.gbif.org/) is one of the biggest platforms in biodiversity, supporting both researchers and citizen scientists. On the other hand, there has been an emphasis on developing institutional repositories in higher education institutions in order to ensure compliance with funder mandates and confirm that data published from the institution meet the necessary standards. Institutional repositories are also useful for cross-disciplinary data where intellectual property might be complicated because of multiple ownership. This type of data will benefit from planning and negotiation services for data acquisition and deposition (Cragin *et al.*, 2010). However, differences in types of research data, data repositories and data sharing policies mean that there are no gold standards for many types of data publishing (Assante *et al.*, 2016).

The potential for future reuse by other researchers is a major incentive for openly sharing research data (Wallis *et al.*, 2013). Therefore, being able to track such reuse is important to understand the value and impact of data and it acts as a reward system for researchers (Costas *et al.*, 2013). While the main focus of data repositories has been on data sharing, information about how and whether data is reused is often not openly accessible from research data repositories. This could be caused by a lack of standard and reliable methods or technological barriers to implement data reuse tracking.

Organizations and initiatives, such as the Research Data Alliance (https://www.rd-alliance.org/), European Data Portal (https://www.europeandataportal.eu/), and FORCE11 (https://www.force11.org/) are developing standard practices and technologies for data support services. Examples include the implementation of persistent identifiers, such as DOIs, for research datasets to aid long-term access and data citation. However, the adoption of such services is not often consistent across all data repositories. It is unclear how the adoption of such data services varies across different types of repositories, what are the challenges that data repository managers face when offering user support and the type of technological solutions that they will benefit from. The aim of this article is to explore the current landscape of data repositories to understand the structure of repository services and types of support offered by them. Furthermore, it examines the current status of the tracking and exposing of data reuse metrics, existing technological barriers and challenges, and type of technologies that may be beneficial in the future.

In order to study different types of data repositories, this article uses Re3data.org, a registry of research data repositories that was established in 2012. It includes a list of data repositories from across the world and publishes information associated with them using the re3data vocabulary. By 2020 the platform listed over 2,500 data repositories, a 6-fold growth in 7 years (Pampel et al., 2013), making it a rich source of information about data repositories globally.

Literature review

Data sharing and use of data repositories

When it comes to storing and sharing electronic research data, until recently the scientific community relied on ad hoc solutions, such as personal storage devices or websites, fulfilling individual data requests by email (Wallis *et al.*, 2013). However, the growth of the open science movement over the past decade has led to the creation of data repositories to provide a reliable infrastructure to preserve data and provide access to it.

A growing body of research has explored the data sharing and reuse behavior of scientists (Federer *et al.*, 2015; Yoon, 2016; Kim and Yoon, 2017; Faniel *et al.*, 2016; Pasquetto *et al.*, 2017). Even

though researchers are often reluctant to openly share their data, the citation advantage of articles that share data can be an incentive (Wallis *et al.*, 2013). However, the use of data repositories and the inclusion of precise data availability information in research outputs are still not commonplace. A survey of biomedical scientific and clinical staff found that 61% had no experience of uploading data to a repository even though they were willing to do so. At the same time, the large number of different types of data repositories means that determining where to upload data can be confusing for researchers (Federer *et al.*, 2015). Many journals now mandate data access statements in articles, but the information provided in these statements can vary. Colavizza *et al.* (2020) compared two open-source journals and found that the percentage of articles that link to a data repository in their data access statements is only 12.2% (6,656 out of 54,719) for BMC and 20.8% (9,013 out of 43,388) for PLOS. Thelwall et al. (2020) investigated 314 primary human genome-wide association (GWAS) articles, with only 13% reporting the location of a complete set of summary GWAS data. This shows that while the use of data repositories is increasing, it is still a minority activity, even when the data is standardized and with high value for sharing. Additionally, library practitioners in the UK had previously reported limited data repository engagement by academic staff and researchers (Pinfield *et al.*, 2014). This may be linked to lacking cultures of data sharing in specific disciplines, but it is important to understand what challenges repository managers are facing in providing these services and whether and how incentives are taken to motivate researchers to adopt standard data sharing practices.

Tracking secondary data reuse

Sharing meaningful data can be time consuming, so researchers often want to know how their shared data is reused (Kratz & Strasser, 2015; Wallis *et al.*, 2013). Data reuse is defined as the retrieval and use of a dataset by someone other than the originator, including the first use of data collected for a community (e.g., astronomy datasets from sky surveys: Pasquetto *et al.*, 2017). Several studies indicate that data reuse is growing with the availability of open data through established data repositories. Examples of data reuse cases and repositories include the UK Data Service (Bishop & Kuula-Luumi, 2017) and the Inter-University Consortium for Political and Social Research (ICPSR) (Faniel *et al.*, 2016) for social science data, the National Heart, Lung, and Blood Institute (NHLBI) data repository for clinical trials data (Coady *et al.*, 2017), and the Global Biodiversity Information Facility (GBIF) for biodiversity and zoology data (Khan *et al.*, 2021).

While data reuse has been of increasing interest to the research community, most research has focused on scientists' data reuse practices rather than technological feasibility of data repositories to track secondary data reuse. Previous research provides useful insights into the factors that are considered important when reusing existing data. For example, Kim and Yoon (2017) found that the availability of data repositories is one of the main factors influencing data reuse at the disciplinary level. Similarly, Faniel and Yakel (2017) report that trust in repositories plays an important role and that data processing, metadata availability, and data selection are important when reusing data. The development of standards and the use of repositories varied between three disciplines studied – social science, archeology and zoology, with archeology lagging the others. Yoon (2016) explored the reasons for data reuse failures and found ease of reuse, understanding data through documentation and lack of support, either from institutions, communities, or individuals (mostly referring data producers) to be the significant factors.

With the rapidly evolving role of data repositories, core requirements of storage and access to data now come with other needs, such as compliance to funder mandates, and proper licensing to ensure

reusability. However, the details of non-core practices vary across different types of repositories. In particular, research data services in higher education institutions lack high level technical services and need proper advisory services for the curation and long-term preservation of active data (Cox *et al.*, 2017). It is therefore important to understand whether repositories now tend to meet data reusers' needs and ensure data and metadata quality. Also, understanding the perceived usefulness of different types of data reuse metrics and any technological barriers to implementing these services is critical for planning purposes.

Data repository surveys

Several surveys have identified best practices and the needs of the data repository community. The Confederation of Open Access Repositories (COAR) received 43 responses to their survey in 2016. Half of the respondents used the same platform for publications and research data and the platforms used varied widely, with DSpace and Dataverse being the most common. Engaging researchers in data sharing, lack of institutional policies, and infrastructure for storage and preservation were the top three challenges mentioned (Shearer & Furtado, 2017). A more recent study by LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries) focused of implementation of FAIR (reference) Data principles in 32 repositories with two surveys– one conducted for repository managers (29 responses) and another for the technical staff (14 responses) (Ivanović *et al.*, 2019). Most (81%) repositories were institutional and 41% of the repositories were based on DSpace and 45% had basic data curation support (brief checking, addition of basic metadata or documentation). The study revealed that the understanding and implementation of FAIR principles are often complicated and not fully met by the respondents. In terms of reuse indicators, Kratz and Strasser (2015) surveyed 247 researchers and 71 data mangers, finding that 85% of researchers and 61% of data managers ranked citations to data as the most important reuse metric. Nevertheless, just over 30% of repositories were tracking citations and only 10% were exposing them.

The surveys conducted so far have investigated important aspects of research data support. Nevertheless, the sample sizes for repository-oriented surveys have been small, with most respondents being from institutional repositories, since participant recruitment methods used mailing lists, personal contacts, social media and circulation within relevant professional networks. In addition, the rapid evolution of the field renders surveys obsolete relatively quickly. In response, this study reports a newer and larger survey that uses openly available metadata from re3data to collect contact information for recruitment purposes, to explore the current technical developments, adoption and change in standard practices, challenges and future needs of data repositories by addressing the following research questions.

RQ1. How do different types of repositories vary in their adoption of technical frameworks? Are additional data support services used by repository managers for data publishing and impact measurement?

RQ2. What kind of data reuse metrics are currently being tracked and exposed by repositories? Are there any technological barriers to collecting these metrics?

RQ3. How does research data support work for different types of repositories and how do they maintain data quality?

RQ4. What are the current challenges and priorities in supporting research data, and what type of tools would be valuable for the future?

Methods

A cross-sectional online survey was selected as the instrument to answer the research questions regarding the current landscape of research data repositories and to compare with the results of previous studies (Fink, 2003). A questionnaire consisting of 13 questions was designed in three main sections – 1) Type of data and technical infrastructure, 2) Research data services and data reuse metrics, and 3) Research data support. These questions were informed by the existing literature and previous data repository surveys (Kratz & Strasser, 2015; Ivanović et al., 2019; Shearer & Furtado, 2017).

Since previous studies focused on a specific type of data repository (Cox et al., 2017) or reported regional distributions of repositories (Shearer & Furtado, 2017), an overview and comparison of research data services from different types of repositories was missing. Therefore, this survey was designed to collect information from four main types of repositories based on the type of data they collect – 1. Institutional repositories, 2. Discipline specific repositories, 3. Cross disciplinary repositories, and 4. Repositories supporting specific data formats only. Regional distribution was not taken into account for this survey since discipline specific and cross-disciplinary repositories are often not limited to a specific country or region.

Questions regarding data reuse metrics were adapted from Kratz & Strasser (2015) to understand the current status and interest in tracking different metrics, with an additional question to capture the different challenges that repositories are facing. Questions regarding research data services were designed around the size of departments, methods used for data quality checking and engagement with users in order to answer RQ3. A multiple-choice question on current challenges was adapted from the findings of Shearer & Furtado (2017). Finally, two open-ended questions were designed to explore current priorities and future tools to advance functionalities of data repositories. Prior to circulating the survey, a pilot study was conducted with the research data librarians at the University of Bath to validate the questions and necessary adjustments were made.

While most previous surveys recruited survey participants via professional channels and often focused on specific countries or regions, this study took a different approach. It attempted to reach more varied data repositories by using the Registry of Research Data Repositories, re3data.org. This registry is based in Germany and began in 2012 but seems to have become a relatively comprehensive source of information about data repositories. It provides a set of relevant metadata about repositories via its application programming interface (API), including contact information. This was selected as the source to collect contact information of repositories where available.

Data Collection

a) Metadata and email addresses from the re3data API

Repository metadata was retrieved from re3data.org on February 23, 2019 from its API. In total 2,274 repositories were listed in the registry at the time of data collection. Its metadata fields included a unique identifier for each repository, name, description, contact, type, start date, end date, language, URL, content type, provider type, keywords, subject, entry date, date the record was last updated and remarks.

An initial inspection indicated that some listed repositories had discontinued service and contact information is not always available since it is not a mandatory metadata field on re3data.org. Furthermore, records had contact information in two different formats – email address or contact form. Since contact forms are not suitable for the survey platform, all records were manually checked for valid email addresses. Data cleaning based on contact email availability and formatting issues resulted in 1,117 records with an email address. When no email was available,

the repository websites were checked instead, finding an additional 70. This resulted in 1,187 curated email addresses for the survey.

b) Survey data
The Jisc Online surveys platform was used to send survey invitations. The survey platform supports two forms of invitations – individual email invitations where the link is valid for a specific recipient and an open survey that can be shared via its URL. The survey opened on June 27, 2019 and closed on October 4, 2019. 1,187 invitations were emailed, with 168 responses (response rate 14%). Survey URLs were also sent out to repositories via contact forms when no email address could be found, and survey URLs were forwarded to previously unknown repositories mentioned by respondents. These methods produced 22 extra responses.

Data Analysis

In total 190 responses were received but 189 were used for analysis after review. One of the responses was excluded from the analysis because of incomplete and premature submission. All responses were anonymized, and any identifiable personal information was removed from them.
The survey questions associated with research questions 1, 2 and 3 were either single or multiple-choice questions with an optional 'Other' field. These answers were analysed to find frequency of responses for different groups and content analysis was conducted where open-text answers were included in the 'Other' field. To explore research question 4, thematic analyses of open-text answers were conducted, and two sets of codes were established by the authors. In total 11 themes were found after reviewing 152 answers for the question regarding current priorities in supporting research data. Similarly, 112 answers for the question regarding future tools and services were reviewed, resulting in 9 themes.
Three coders independently reviewed and coded the responses for each theme – 1 if a response corresponds to a theme or 0 otherwise. This was considered a complicated task because of variations in length and wording of responses. A Fleiss's kappa test was conducted to calculate interrater reliability (table 2 and 3). There was moderate (0.41-0.60) to substantial agreement (0.61-0.80) in most cases. Where kappa values were below 0.4 (table 2), most of the votes fell into one category (code 0 for negative in this case) with a low-level of agreement for the rest. Thus, kappa values were low despite the high level of agreement for a single category. Obtaining high kappa values is difficult for very unbalanced classification tasks, explaining the low agreement rates (Hripcsak & Heitjan, 2002). Disagreements between coders were resolved after a discussion among the coders. These values were reported as the number and percentage of responses in table 2 and 3.

Findings

Type of data repositories and technical frameworks

In total 189 responses were received from data repository librarians or data managers, with a majority of responses from institutional and discipline-specific repositories. A high percentage of repositories (71%) used technical frameworks other than DSpace, Eprint and Fedora (Table I), although it is possible that repositories not responding tended to use a standard framework but did not know its name to report. Other types of frameworks included bespoke systems developed in house, Dataverse and Figshare for Institutions. Bespoke solutions include custom built systems written with combinations of Comprehensive Knowledge Archive Network (CKAN), Ruby on Rails, Socrata, SQL, Java, XML web application, Solr, Mongo, Dojo, Python, and MySQL. Some

repository developers used other software systems, such as Invenio (open-source software to build large-scale information systems - https://inveniosoftware.org/), Islandora (open-source digital asset management system - https://islandora.ca/), Archivematica (open-source digital preservation system - https://www.archivematica.org/en/), LibreCat (institutional repository system – https://github.com/LibreCat/LibreCat), and Adobe Coldfusion (commercial rapid web-application development computing platform).

Institutional repositories are more likely to use DSpace, Fedora, Eprint, Dataverse, and FigShare for Institutions. Perhaps existing repository frameworks are relatively easy to adopt, and this helps academic institutions develop repository services promptly, often without specialist technical support.

Table I. Type of data repositories and technical frameworks used

Type of data collected	*Responses*	*Percentage (%)*	*Type of repository frameworks*
Institutional	64	34%	16% Dspace, 12% Eprint, 3% Fedora, 66% other types.
Discipline specific	89	47%	8% Dspace, 3% Fedora, 73% other types.
Any disciplines	22	12%	14% Dspace, 5% Eprint, 77% other types.
Specific data format	4	2%	100% other types.
Other	10	5%	20% Fedora, 60% other types.

Most repository services supported use of persistent identifiers (PID) for datasets with only 8% not supporting any PIDs, where some supported a combination of PIDs. 76% repository services supported DOI, 22% supported Handle (http://www.handle.net/), and 21% supported other types of identifiers, often specific to a data type or discipline. In addition, 48% repositories used DataCite (https://datacite.org/) as a DOI provider, 7% used the Data Citation Index to track data citations and 5% used other data services, such as IRUS-UK (Institutional Repository Usage Statistics UK), Altmetrics, Scholix (http://www.scholix.org/).

Data reuse metrics

Repository managers were asked which of the following data usage metrics would be helpful, whether they currently collect them or not: citation counts, download counts, landing page views, and links to the data from other websites (e.g., educational use, Wikipedia) (Figure 1). Overall, 57% of respondents considered citation counts extremely useful and 30% considered them very useful. Among different types of repositories, 61% of institutional and discipline specific repository managers and 41% of repository managers who collect data from any disciplines considered them extremely useful. Download counts and links to the data from external websites were considered very useful metrics by nearly half: 41% and 44% respectively. Landing page views were less valued, with 28% considering them very useful and 29% moderately useful. This is in line with the findings of Kratz and Strasser (2015) except there has been a growing interest in links to data from other websites.

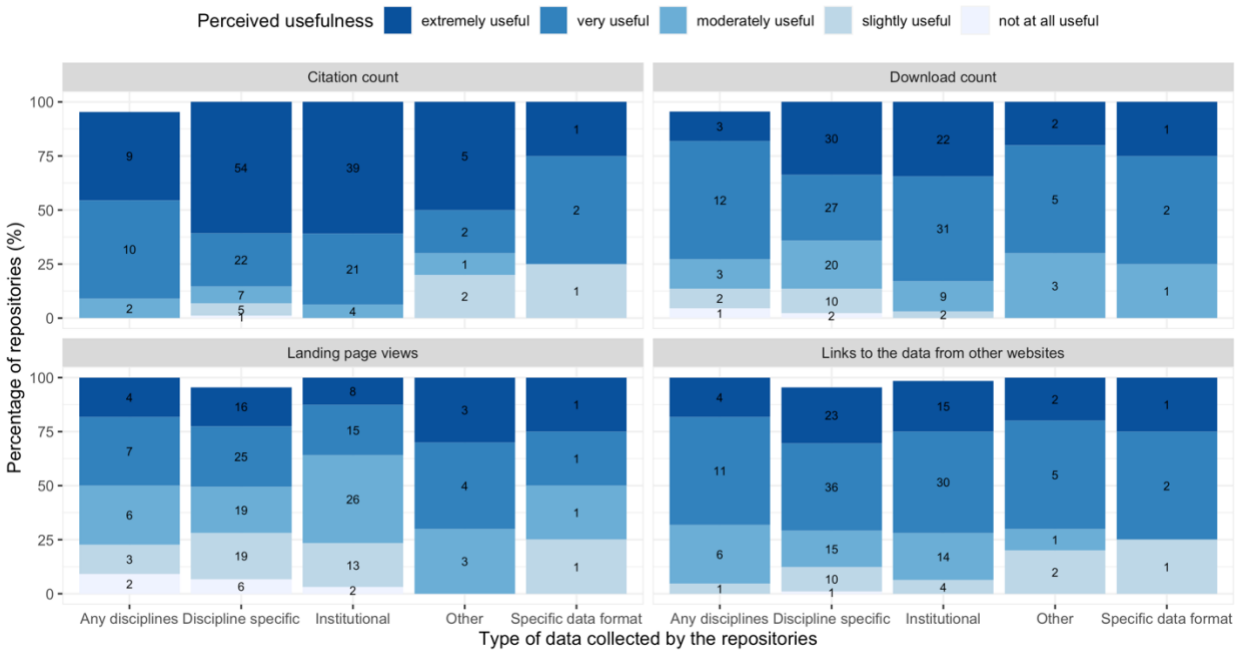

Figure 1. Perceived usefulness of different type of metrics (labels on bars represent number of responses)

Overall, 50% repository managers responded that they currently do not track secondary reuse cases of their published datasets but interested, 32% mentioned they currently track data reuse cases in some format, and 19% were not interested in doing so. Among these groups, 38% discipline specific, 25% institutional, 18% cross-disciplinary, 25% of specific data format supporting repositories and 50% of other repository managers currently track data reuse metrics, but further 64% institutional, 49% discipline specific and 41% cross-disciplinary repository managers are interested to implement this in the future.

A follow-up question was asked to 32% repository managers who currently track data reuse – tracking and display status of specific metrics, and barriers to tracking these metrics in cases these are untracked. Figure 2 shows the tracking status of four different metrics by different repositories – download counts, citation counts, views, and citations to the repository. Similar to the findings of Kratz and Strasser (2015), downloads and views are more frequently tracked by repository managers, followed by citations to datasets and citations to the repository as a whole. Among those who track these metrics, few tend to expose them on their platform (Figure 3).

Dataset citations

23% of cross-disciplinary (n=5), 46% of discipline specific (n=41), 33% of institutional (n=21), 25% of specific data format (n=1), and 50% of repository managers in other groups (n=5) reported that they track citations to datasets (Figure 2). Within these groups, all cross-disciplinary repository managers and nearly half of the institutional and other repository managers display this metric. The percentage is slightly lower (39%) for discipline specific repositories and the repository manager in the specific data format group did not respond (Figure 3).

Repository managers reported the following reasons for not being able to track or expose dataset citations: difficult to enforce and track dataset citations as research articles often do not include

dataset citations in their main references; Data Citation Index does not harvest data related to all repositories; lack of reliable technologies to automate the process, resulting in users having to manually report any citations. Some repository managers reported using Google Scholar, euroPMC and Dimensions as a source of citation data, but lack of trust and reliability in citation data may have resulted in less repositories exposing the results.

Downloads

Most repository managers (80-100%) in all groups reported tracking downloads (Figure 2). However, among these groups, 50% of cross-disciplinary (n=9), 73% of discipline specific (n=53), 29% of institutional (n=15), 50% of specific data format (n=2) and 80% of other repository managers (n=8) do not expose download counts in their repositories (Figure 3). Besides technical difficulties and privacy concerns, lack of interest was mentioned as an important factor in displaying download counts. Some repository managers offer this service only internally. One participant mentioned concerns about the reactions of researchers to these numbers. Another participant raised the technical concern that sections of datasets were often downloaded instead of entire datasets, so a download count for whole datasets were not meaningful.

Figure 2. Tracking status of different type of metrics (labels on bars represent number of responses)

Views

Similar to download counts, 70-100% of repository managers for different types of repositories mentioned tracking views, even though most of them do not expose the numbers (Figure 2 and 3). While some repository managers share view counts internally, many mentioned that page views are of less interest to stakeholders as this metric is not significant and can be manipulated easily. A few participants mentioned privacy issues such as disabling the tracking of metadata discovery and views because of GDPR regulations and distrust in web-trackers.

Citations to the repository

This metric was the least tracked by all types of repository managers. Whilst 40% (n=36) of discipline specific repositories mentioned tracking this, 39% (n=14) of them exposed this metric. Most of the repository managers did not consider this a valuable metric compared to citations to individual datasets and found it difficult to technologically implement.

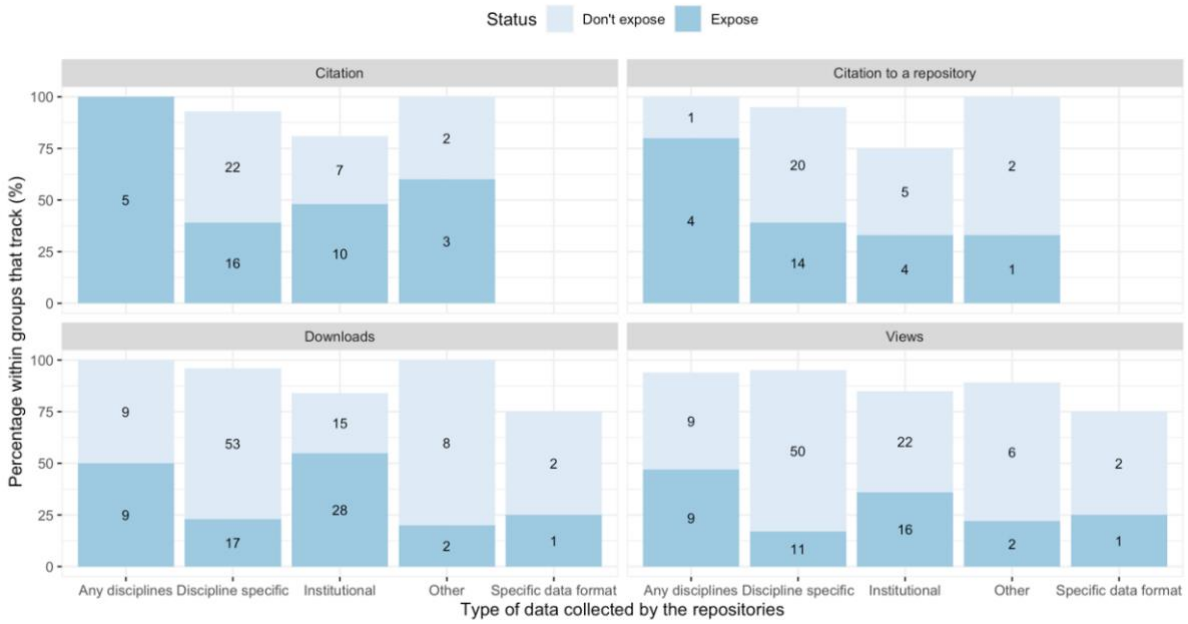

Figure 3. Metrics display status of repositories that track them (labels on bars represent number of responses)

Research data services and quality maintenance

Overall, 34% of research data services run as small departments of two or three members. Among the rest, 25% are larger departments with more than three people, 19% are solo services, 6% provide no institutional support and 15% mentioned other approaches, such as spreading services over multiple departments without a designated research data service department. Figure 4 shows the distribution of types of service for different repositories.

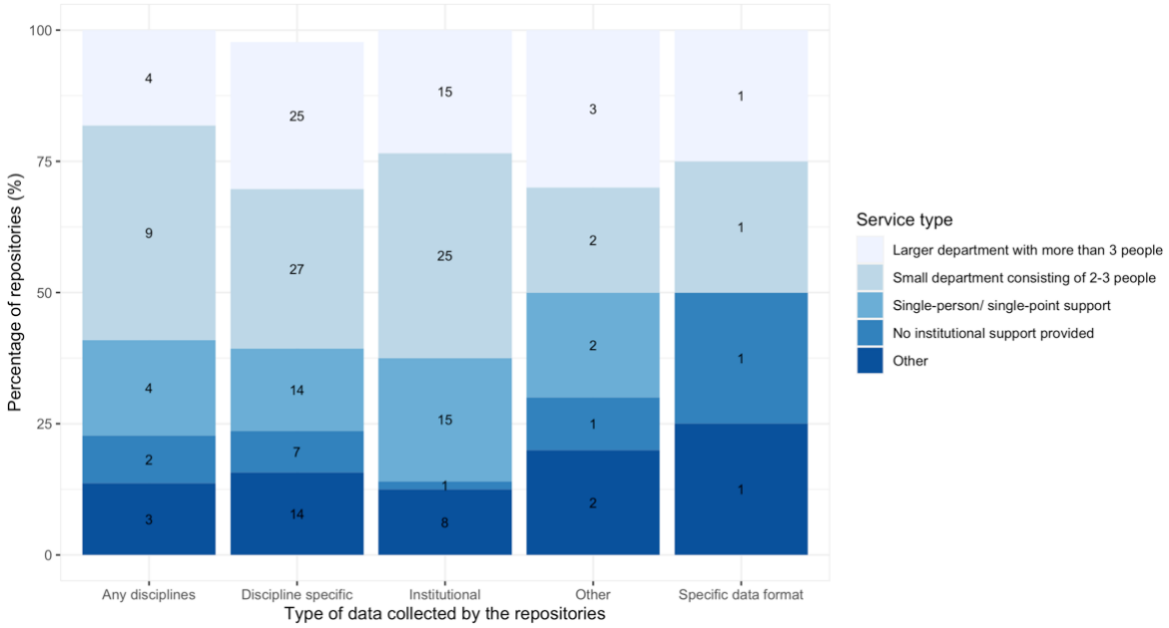

Figure 4. Type of research data services provided by the repositories

Most repository managers support dataset and metadata quality checks via librarians, subject specialists or information professionals (Figure 5). This was similar across different types of repositories and research data support services, except where no institutional support was provided – automated checks were more frequently used in those cases. Where participants mentioned other types of quality maintenance methods, most combined automated checks, (e.g., using scripts to look for errors and duplication) and manual checks by a designated member (e.g., librarian).

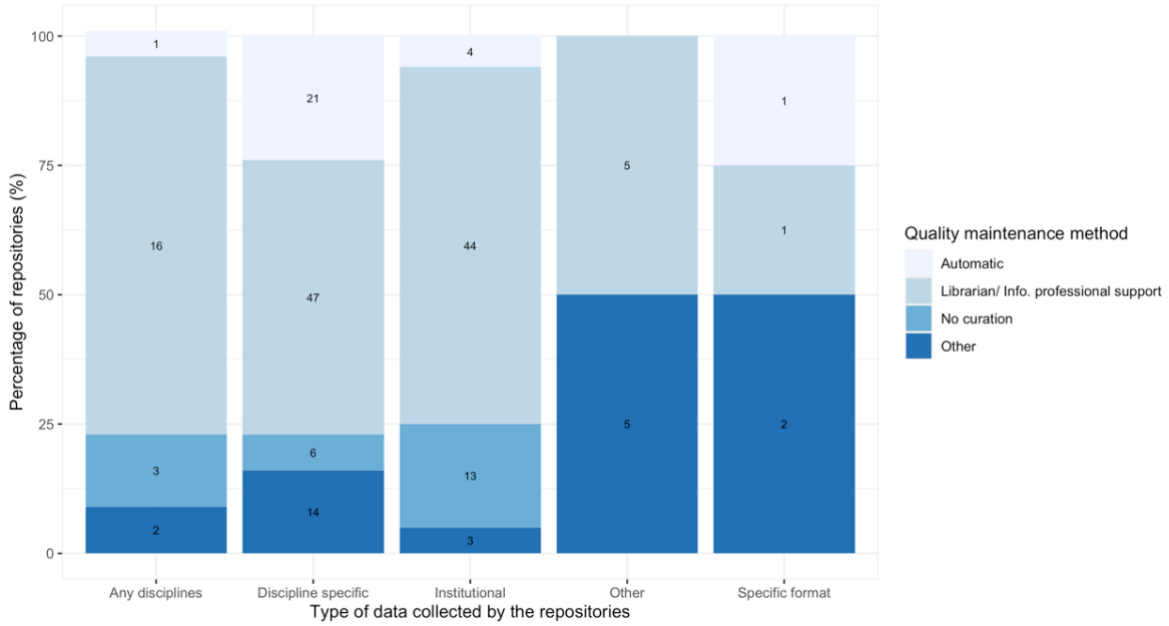

Figure 5. Data quality maintenance types by the repositories

Challenges and motivators

Lack of engagement from the users and lack of human resources are the top two challenges mentioned by repositories across all groups (Figure 6). 73% of institutional, 64% of cross-disciplinary, 50% of specific data format and 40% of discipline specific repository managers mentioned lack of engagement to be a challenge. Long-term maintenance was also a major concern among all repositories, whereas a lack of user need was mentioned by fewer repositories. Lack of funding was mentioned as a challenge by nearly half of the discipline specific repositories. Other user challenges include researchers' lack of understanding of standards requirements, multiple user defined data protocols, trends to put resources in multiple websites, and diverse user needs. Identifying standards, legal or data ownership issues and deciding a long-term solution in an evolving field were some service challenges mentioned by the participants. Adding new functionalities to existing systems, improving metadata quality and assessing the quality of datasets for publishing without domain expertise are also challenging issues. One participant mentioned demonstrating the value of published data and being able to integrate any technological methods on top of current repository systems -

"Tracking usage of data to demonstrate value of the repository […]. We have minimal resourcing to better implement solutions that do exist."

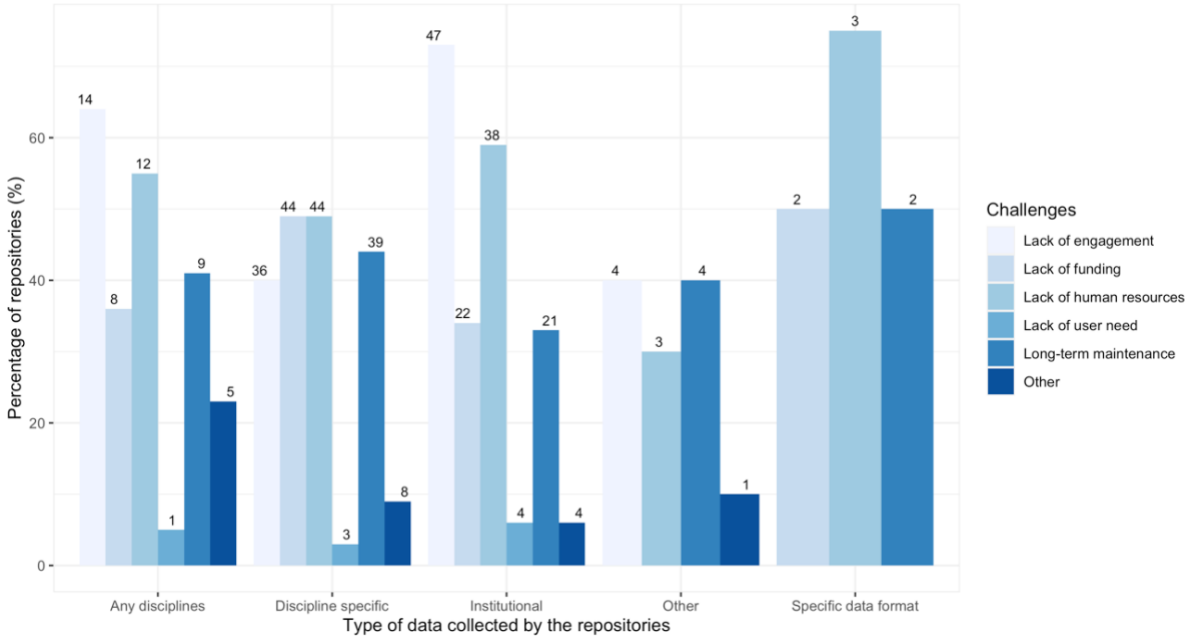

Figure 6. Type of challenges faced by the repositories

On average, outreach was mentioned as the most common way to motivate researchers by different repositories (69%,). This was followed by funder policies (59%) and training programs (56%). Funder policies was a significant motivating factor for academic researchers who use institutional repositories (77%), compared to 54% discipline specific and 50% cross-disciplinary repositories (Figure 5). 23% responses selected other, which includes journal mandates, developing innovative programs, such as Research Data Champions, research data management policy, and utilizing different channels of communication. Several mentioned no active input to motivate researchers as the repositories are well established and used by researchers according to their needs.

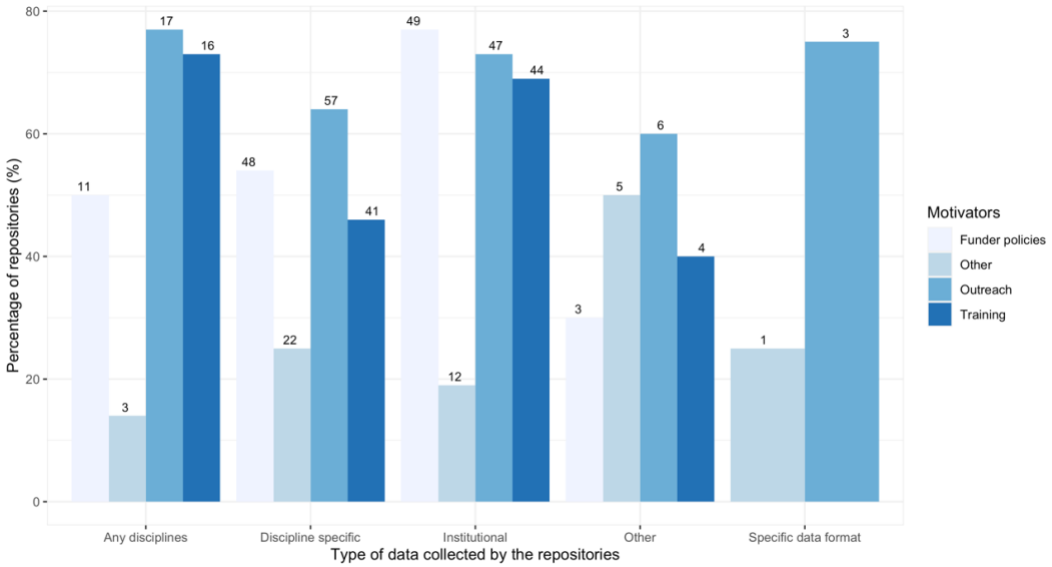

Figure 7. Motivators for researchers to deposit data in repositories

Priorities

In total 152 responses were received for the open-ended question regarding current priorities in research data support. These responses are grouped under 10 categories with each response fitting one or more (Table II). 49% of respondents mentioned ensuring that data is FAIR (Findable, Accessible, Interoperable and Reusable) as a top priority (Wilkinson et al., 2016). Other high priorities include providing user support for research, e.g., data management plan (DMP) review (36%) as well as building relationships and developing best practices, e.g., provide research data management (RDM) training (29%). As repositories are handling an increased volume of data, ensuring data and metadata quality (15%), simplified data handling (10%), and building robust infrastructures with improved search systems and other data management features (15%) are also considered important. With growing needs from users, some repositories are considering inclusion of other data support services to support data usage and analysis (6%). The following response demonstrates how data repositories are dealing with a multitude of challenges and having to set priorities accordingly –

"Educating those gathering data about improving data management practices, e.g., FAIR Principles (New Zealand is very behind on this); improving application of data management practices among scientists; providing simple to use DM tools; managing data privacy issues, e.g., for data collected on private land; dealing with the challenges of big data and data science, e.g., data volumes; managing data and metadata where Edge computing and sensors are being used; provenance, repeatability and fine grained metadata"

Table II. Main priorities for repositories (n=152)

Type of priorities	*Number of responses*	*Percentage (%)*	*Agreement rate (%)*	*Kappa value*
FAIR data	75	49	68.4	0.58
User support (DMP)/ research support)	54	36	75.7	0.61

Outreach, RDM training, build relationships, and develop best practices	44	29	91.2	0.51
Data quality check	22	15	69.7	0.49
Robust infrastructure (improved search system, development and inclusion of new data management features)	22	15	84.2	0.55
Simplified data handling for ease of use	15	10	82.2	0.36
Need for a systematic approach (norms/ standards/ compliance)	14	9	83.6	0.18
Support data services (e.g., support data access, use, analysis etc.)	9	6	90.1	0.45
Secure funding	6	4	90.1	0.18
Better usage metrics	4	3	95.4	0.35
Inclusion of data access statement	2	1	95.4	0.45

Tools needed in the future

114 participants responded to the open-ended question regarding the type of tools or research data support system they envision for the future. These responses were grouped into nine categories (Table III). Among different recommendations that emerged, integration and interoperability between data and systems was considered important by the most (30%). One participant viewed convergence of data, publications and research intelligence functions as the ultimate solution to move forward since current systems are very distributed and non-communicating. Other participants mentioned integration of internal institutional systems, such as a Current Research Information System (CRIS), a DMP tool, repositories to allow reuse of metadata, as well as integration with specialized services (e.g., visualization, data aggregation) on top of their archived data.

Improved research data management tools (e.g., machine readable DMP) as well as building community of practice across country and developing and sharing more training material was suggested by 19% of the respondents. For example, a DMP wizard would be useful that gives the researchers all features and issues to consider when starting a data production or packaging a project. 16% participants recommended automated systems for data handling, linkage between publications and datasets, and metrics tracking, and tools that could allow performing any data analysis and visualization without downloading individual datasets.

In terms of adopting a repository service, a national infrastructure or federated repository was recommended by 15% that would allow a simpler local setup and shift their emphasis other new data services and features. Similarly, better data processing, discoverability and storage for repository systems were recommended by some participants, such as tools for long-term preservation of data, streamlined PID based systems, improved search systems, and integration of tools to capture and manage data, e.g., electronic lab notebooks, Open Science Framework (OSF) for research workflow.

Table III. Type of tools and services needed in the future (n=114)

Type of tools/ services	*Number of responses*	*Percentage (%)*	*Agreement rate (%)*	*Kappa value*
Integration and interoperability between data and systems (e.g., data exchange between different	34	30	87.7	0.77

domains/ journals and repositories via national framework, federated systems, ontology tools)				
Better RDM tools (enhanced DMP Wizard, machine readable DMP), promote standardization, develop community practices across countries	22	19	79.8	0.49
Tools that allow computation (e.g., analysis and visualization of data) without downloading datasets	18	16	82.5	0.54
Automated systems (data identification, quality check, import/export of data/metadata, linking between publication and data, metrics tracking)	18	16	87.7	0.67
Repository framework with simpler local setup with emphasis on visualization and analytical service; APIs; new features	17	15	80.7	0.44
Tools supporting long-term preservation of large volume of data	8	7	87.7	0.49
Streamlined persistent identifier (PID) based systems with better handling of versions and subsets	6	5	95.6	0.69
Powerful search engine for better data discoverability	6	5	91.2	0.51
Integration of tools to capture and manage data (e.g., lab instruments, OSF, Electronic lab notebooks, Sandbox)	5	4	93.9	0.67

Limitations

The sample size analysed here is small given that over 2000 research data repositories are indexed on re3data.org. This shortfall is partly because of the limited availability of email addresses provided for repositories in the metadata. Where web forms are used as a contact method, automated distribution of the survey via the platform is impossible. In addition, email addresses provided by repositories tended to be generic email addresses, so in many cases the email invitation to the survey had to be forwarded to the repository manager to answer the questions. This indirect route slowed down the process and reduced the number of responses. This may also bias the sample against non-English repositories (less likely to forward an English email), and small repositories (without a dedicated contact email). The sample sizes for certain repository types, such as repositories supporting specific data formats, is relatively small and the results may not accurately represent that group.

Recommendations for registries of data repositories

This study used openly available metadata from re3data.org as a source of contact information to recruit participants. Registries of data repositories, such as re3data.org, and general repositories, such as OpenDOAR (https://v2.sherpa.ac.uk/opendoar/), are valuable for similar studies in the future but can benefit from more granular and structured metadata in some data fields. We recommend the following changes to help with this. Of course, these suggestions are secondary to the primary purpose of registries, which is to collect and index the world's repositories.

For the optional 'Contact' field in re3data and OpenDOAR, we recommend defining contact types into two main groups – 1. Email address, 2. Web address/contact form, where the email address field should accept valid email addresses only. Additionally, we recommend using more granular 'Repository types' to accommodate differences between federated infrastructures, data portals and data repository, and also differentiate between other repository types, such as cross-disciplinary, government, project-based repository. Mandating the optional Software field that is currently

included by 12% repositories only would allow future studies to analyse software-based differences, which may be important to repositories. Similarly, currently only 0.6% repositories complete the 'Quality management' field where it accepts three answers – yes, no and unknown. Since data quality is one of the main factors that drives future data reuse, it would be beneficial to further extend this field based on existing research to provide more information on the type of quality management implemented. This will benefit researchers searching for data repositories to deposit data or to find existing datasets.

Discussion

Data repositories are evolving rapidly to accommodate the needs of funding bodies and researchers, and to support different types of data. Because of differences in the nature of disciplinary data, most discipline-specific repositories developing bespoke technical frameworks. In contrast, a major incentive for institutional repositories is to support academic researchers following funders' policies, but they often lack the technical expertise available to large-scale discipline specific repositories (Cox *et al.* 2017) and rely on existing frameworks, such as DSpace, Eprint, and Dataverse. Shearer and Furtado (2017) and Ivanović *et al.* (2019) found similar results for technical framework adoption by institutional repositories. While differences in institutions and data types mean a single repository framework will not fit all purposes, there are opportunities to use community driven approaches for developing research data management policies, training materials and best practices.

As indicated by this study, lack of engagement and lack of human resources are the major challenges faced by all repositories, but these issues were more prominent for institutional repositories than for discipline-specific repositories within the study sample. Shearer and Furtado (2017) also reported lack of engagement as the top challenge among their respondents. Comparatively higher percentages of institutional repositories therefore heavily rely on outreach and training to motivate and engage researchers, even though funder mandates were the main motivators for academic researchers who use these repositories. Outreach, RDM training, building relationships, and developing best practices was mentioned among the top three priorities in this study, and institutional repositories can benefit from working together to develop training materials and policies where a lack of human resources is an operational issue.

Most research data repositories started as siloed services to give access to data. As these services mature, better data discoverability and interoperability are becoming increasingly important to promote data reuse. This reflected in the findings of this survey, because integration and interoperability between data and systems was the most wished for future service. There are two different routes to achieve this – 1. Interoperability between data repositories and 2. Interoperability between journal systems and data repositories. The first route requires a data portal or a global discovery service that would break the silo of individual repositories and allow users to search data across multiple repositories. Federated infrastructures achieve this by connecting multiple data repositories and act as an access point for data across those repositories. Some disciplines, such as biodiversity (GBIF.org), and national initiatives, such as the National Research Data Infrastructure (NFDI) for the European Open Science, show that this can be possible (Chamanara *et al.*, 2019; Goldstein, 2017). Another relatively new data discovery service is Google Dataset Search (https://datasetsearch.research.google.com/), which allows dataset keyword searches across all supported repositories. This system is based on a linked data model and relies on repository services adopting the schema.org metadata standard so that dataset

metadata from these repositories can be indexed by Google and added to the search system (Patel, 2019).

Interoperability between journal systems and data repositories is necessary to automate tracking data citations to understand how a published dataset has been reused. We found that repository managers considered dataset citations as the standard metric to estimate the scholarly value of data, as well as valued evidence of educational use. Download counts and views are considered less valuable since they are not evidence of secondary use and can be manipulated easily. Despite their importance, there is no standard method to count dataset citations that can be implemented across different repository systems. Initiatives and technical frameworks, such as Scholix (Scholarly Link Exchange) are currently in progress. Khan *et al.* (2020) suggest further enhancements of the Scholix schema and enrichment of Scholexplorer metadata using controlled vocabularies and the adoption of standardized data citations by journals to establish links between datasets and literature. Google Dataset Search also displays citation counts for datasets, but Khan et al. (2021) found discrepancies between these numbers and the citation counts displayed by Google and by GBIF. These services can be potentially used to identify data reuse cases when they mature in the future. In the meantime, repositories should follow and implement the data citation roadmap (Fenner et al., 2019), and carefully consider the guidelines for using indicators to evaluate data outlined by Konkiel (2020).

Conclusion

This study identified the key current practices of data repositories and the types of challenges they face. Our results show that the sporadic development of different types of data repositories has resulted in the adoption of bespoke technical frameworks by the majority, and especially by discipline-specific repositories. However, developing and implementing new technological solutions for different platforms can be challenging for institutional repository services as we found that they often had small teams. Whilst it seems logical that disciplinary repositories would often need bespoke services, this makes full interoperability between services difficult to achieve. Nevertheless, integration and interoperability between data and systems was considered important by the respondents. A common language that can be used by all repository systems can help break this silo, such as the Schema.org metadata standard for datasets to be indexed and discovered by Google Dataset Search, and adherence to standard data citation practices by both researchers and journals. Additionally, this will also help repository services track and expose data reuse metrics, such as citation counts for datasets, as suggested by the survey results.

In the long-term, the use of federated systems and simpler local set-ups will allow repository services to shift their focus to new features, such as the integration of tools to capture and manage data (e.g., lab instruments, Open Science Framework (OSF), Electronic lab notebooks, Sandbox) and the development of new visualization and analytical tools. Where majority of repository services are currently struggling with a lack of user engagement, these new improvements will help demonstrate the value of research data and attract more users.

Given the apparent mismatch between the features desired by repositories and the availability of large enough teams to implement them, current collaborative initiatives seem likely to help to develop shared community practices and reduce the burden on individual institutions. Global initiatives, such as the implementation of FAIR data practices, Scholix and Google Dataset Search will benefit all repository types by promoting standardization, improved data discoverability and the automation of secondary data reuse tracking. While institutional policies and types of outreach activities to engage researchers can differ between academic institutions, shared resources to

implement technological solutions (e.g., how to use Schema.org metadata standard for a specific repository framework), guidelines and training materials for research data management will be helpful, especially for smaller scale academic institutions. This will also ensure that different data repositories are not developing siloed services but have a common interoperable system in place. Data sharing and data protection rules can vary across different countries and regions, such as Europe has General Data Protection Regulation (GDPR) in place and some survey participants mentioned this as a barrier to exposing certain data metrics. Regional collaboration will be valuable in these cases.

Conflicts of interest/Competing interests

Not applicable

Data availability statement

Anonymized survey data from this research is available at https://doi.org/10.6084/m9.figshare.14191739.

References

1. Assante, M., Candela, L., Castelli, D., & Tani, A. (2016), "Are scientific data repositories coping with research data publishing?", *Data Science Journal*, Vol. 15. http://doi.org/10.5334/dsj-2016-006
2. Bishop, L., & Kuula-Luumi, A. (2017), "Revisiting qualitative data reuse: A decade on", *Sage Open*, Vol. 7 No. 1, 2158244016685136.
3. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M. and Schindler, U. (2017), "The Scholix framework for interoperability in data-literature information exchange", *D-Lib Magazine*, Vol. 23 No. 1/2.
4. Chamanara, J., Kraft, A., Auer, S., & Koepler, O. (2019), "Towards Semantic Integration of Federated Research Data", *Datenbank Spektrum* Vol. 19, pp. 87–94. https://doi.org/10.1007/s13222-019-00315-w
5. Coady, S. A., Mensah, G. A., Wagner, E. L., Goldfarb, M. E., Hitchcock, D. M., & Giffen, C. A. (2017), "Use of the national heart, lung, and blood institute data repository", *New England Journal of Medicine*, Vol. 376 No. 19, pp. 1849-1858.
6. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020), "The citation advantage of linking publications to research data", *PloS one*, Vol. 15 No. 4, e0230416.
7. Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013), "The value of research data–Metrics for datasets from a cultural and technical point of view", *A Knowledge Exchange Report*, Doctoral dissertation, Leiden.
8. Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017), "Developments in research data management in academic libraries: Towards an understanding of research data service maturity", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 9, pp. 2182-2200.
9. Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 368 No. 1926, pp. 4023-4038.

10. Faniel, I. M., Kriesberg, A., & Yakel, E. (2016), "Social scientists' satisfaction with data reuse", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 6, pp. 1404-1416.
11. Faniel, I. M., & Yakel, E. (2017), "Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation", *Curating research data, volume one: Practical strategies for your digital repository*, 1, pp.103-126.
12. Federer L. M., Lu Y-L, Joubert D. J., Welsh J., Brandys B. (2015) "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff", *PLoS ONE*, Vol. 10 No. 6, e0129506. https://doi.org/10.1371/journal.pone.0129506
13. Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M. and Clark, T. (2019), "A data citation roadmap for scholarly data repositories", *Scientific Data*, Vol. 6 No. 1, pp. 1-9.
14. Fink, A. (2003), *How to design survey studies*, Sage.
15. Goldstein, S. (2017), "The Evolving Landscape Of Federated Research Data Infrastructures", available at: https://www.rd-alliance.org/sites/default/files/attachment/The_Evolving_Landscape_of_Federated_Research_Data_Infrastructures.pdf
16. Hripcsak, G., & Heitjan, D. F. (2002), "Measuring agreement in medical informatics reliability studies", *Journal of biomedical informatics*, Vol. 35 No. 2, pp. 99-110.
17. Ivanović, D., Schmidt, B., Grim, R., Dunning, A. (2019), "FAIRness of Repositories & Their Data: A Report from LIBER's Research Data Management Working Group", available at https:// doi.org/10.5281/zenodo.3251593
18. Khan, N., Thelwall, M. and Kousha, K. (2021), "Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics", *Scientometrics*, pp.1-19.
19. Khan, N., Pink, C. J., & Thelwall, M. (2020), "Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations", *Patterns*, Vol. 1 No. 1, 100007.
20. Kim, Y., & Yoon, A. (2017), "Scientists' data reuse behaviors: A multilevel analysis", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 12, pp. 2709-2719.
21. Konkiel, S. (2020), "Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators", *Scholarly Assessment Reports*, Vol. 2 No. 1.
22. Kratz, J. E., & Strasser, C. (2015), "Making data count", *Scientific Data*, Vol. 2 No. 1, pp. 1-5.
23. Pinfield, S., Cox, A. M., & Smith, J. (2014), "Research data management and libraries: Relationships, activities, drivers and influences", *PLoS One*, Vol. 9 No. 12, e114734.
24. Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.J., Gundlach, J., Schirmbacher, P. and Dierolf, U. (2013), "Making research data repositories visible: the re3data. org registry", *PloS One*, Vol. 8(11), e78080.
25. Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017), "On the reuse of scientific data", *Data Science Journal*, Vol. 16, 8.

26. Patel, D. (2019), "How Google's Dataset Search Engine Work", available at: https://towardsdatascience.com/how-googles-dataset-search-engine-work-928fa5237787 (accessed 31 March 2021)
27. Shearer, K., & Furtado, F. (2017), "COAR Survey of Research Data Management: Results", available at: https://www.coar-repositories.org/files/COAR-RDM-Survey-Jan-2017.pdf
28. "re3data.org - Registry of Research Data Repositories", available at: https://doi.org/10.17616/R3D (accessed 17 November 2020).
29. Robinson-García, N., Jiménez-Contreras, E. and Torres-Salinas, D. (2016), "Analyzing data citation practices using the data citation index", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 12, pp.2964-2975.
30. Thelwall, M., Munafò, M., Mas-Bleda, A., Stuart, E., Makita, M., Weigert, V., Keene, C., Khan, N., Drax, K. and Kousha, K. (2020), "Is useful research data usually shared? An investigation of genome-wide association study summary statistics", *Plos One*, Vol. 15 No. 2, e0229578. https://doi.org/10.1371/journal.pone.0229578
31. Wallis, J. C., Rolando, E., & Borgman, C. L. (2013), "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology", *PloS One*, Vol. 8 No. 7, e67332.
32. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3 No. 1, pp. 1-9.
33. Yoon, A. (2016), "Red flags in data: Learning from failed data reuse experiences", *Proceedings of the Association for Information Science and Technology*, Vol. 53 No. 1, pp. 1-6.