ZBW *Publikationsarchiv*

Publikationen von Beschäftigten der ZBW – Leibniz-Informationszentrum Wirtschaft
Publications by ZBW – Leibniz Information Centre for Economics staff members

Limani, Fidan; Hajra, Arben; Mexhid, Ferati; Radevski, Vladimir

Conference Paper — **Published Version**

Requirements and Recommendations for University Research Data Repository: A Case Study

This Version is available at:
http://hdl.handle.net/11108/514

Kontakt/Contact
ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der
Leibniz
Leibniz-Gemeinschaft

REQUIREMENTS AND RECOMMENDATIONS FOR UNIVERSITY RESEARCH DATA REPOSITORY: A CASE STUDY

Fidan Limani[1], Arben Hajra[1], Mexhid Ferati[2] and Vladimir Radevski[3]

[1]Leibniz Information Centre for Economics, Kiel/Hamburg, Germany
[2]Linnaeus University, Kalmar, Sweden
[3]Southeast European University, Skopje/Tetovo, North Macedonia

ABSTRACT

The changing research culture poses new requirements on research communities across disciplines. The availability of different research outputs requires appropriate infrastructure in order to publish and recognize them as research contributions. Institutional repositories (IR) provide such support for research outputs generated as part of the research. In this work, we conduct a survey to assess the research data management activities and services for an IR at a university setting. The survey results show an increasing presence of research data (RD) as well as a need for an IR component to support it for the participating communities. The survey reveals that common practices among researchers are to save data into spreadsheets, text documents, and relational databases, which they manage personally without any institutional support. Most of them store this data on a personal computer or portable storage drives. Despite these storing practices, which are not optimal for sharing, most participants have a need of sharing their data, and they do it mostly using email attachments. The survey also reveals features that repositories should provide, mostly RD creation- and dissemination-related activities, but also covering those that relate to backup and access-granting.

1. INTRODUCTION

Research undertakings typically generate multiple deliverables, such as research data (RD), source code, workflow models, and so on. Considering this, basing the research dissemination on publications alone is simply insufficient to capture all the research aspects. Moreover, this also hampers research reuse, verification and reproducibility. As the role of the different research deliverables increases, so do research practice expectations to make them available for the research communities.

One such research output that is gaining in importance in scholarly communication, and one that we focus on in our work, is research data (RD). Even though its practice – both in creation and reuse – differs with research communities (for e.g., computationally-driven fields lead in RD generation and reliance on it; see for example Tony (2016) on RD's importance), we are generally witnessing an increase in RD's role in research. This is already obvious with the multiple publication venues that target RD, as is the case with the data paper, the case of RD published alongside publications, or packaged with other research artifacts – source code, publication, funding details, etc. This provides RD with an appropriate recognition – commonly attributed to research publications – which makes for a good motivation for researchers to put adequate effort to manage RD from its creation, through its reuse, and beyond.

Different factors shape the evolving RD research practices across disciplines. Publishers, driven from more research transparency and reuse, more and more require that researchers provide the accompanying RD of the research publications (Lin and Strasser, 2014). On the other hand, initiatives like DataCite, provide the necessary means (metadata standard and services that revolve around RD) to recognize the effort that researchers spend in RD generation ("Welcome to DataCite", n.d.); RD citation and usage or impact metrics present additional incentives to researchers for sharing their RD. Moreover, (public) funding agencies, another important actor in the research ecosystem, have accommodated new RD requirements for grant holders to make

research outputs – and some specifically refer to RD – available. For their Horizon 2020 projects, the European Commission requires a certain level of alignment with more open access-type form of RD publication. Research Data Pilot project (H2020 Programme, 2017) is one such example. Among other things, by being readily available, it is expected that duplication efforts (re)create the same or similar RD decrease.

In the wake of ever-increasing importance of RD in research, a set of standards for making them more usable – not only by humans, but also machines – is emerging. Wilkinson et al. (2016) propose a set of guidelines that would support both humans and machines locate and reuse the different research artifacts. While just a set of guidelines, open to interpretation to different communities, FAIR (Findability, Accessibility, Interoperability, Reusability) Data Principles ("FAIR principles", n.d.) provide a good basis for better-suited RD research practices.

In this context, research communities need the necessary means to handle RD archival and dissemination. It is clear that an infrastructure – repository at an institutional level – is required in order to support these research expectations, while considering the lifecycle and specific domain practices in the community. Lynch (2003), for example, supports the role of such repositories in the scholarly infrastructure.

Defining IRs depends on the context, research practices, domain, and so on. For the purpose of this paper, we adopt the definition from Luther (2018), as an entity that "contains digital materials created by the institution and its community members". The digital materials targeted for preservation and disseminating include research outcomes that stem from or are used in both research and teaching.

Universities are part of the research community and regularly deal, to different extents, with RD. As a result, there is a need for a corresponding support for RD management for the communities. The idea of a repository for any type of resource (publications, research data, workflows, algorithms, etc.) is enticing and generally accepted as advantageous to individual researchers and institutions alike. However, having in mind the different criteria that affect such a project, we need to identify the requirements that researchers have for an RD repository.

To investigate this matter, we look at Southeast European University (SEEU) in North Macedonia as a case study. We survey faculty of the university, and our aim is to investigate the extent that RD and RD-based activities are present at the university and what are the specific needs of the faculty in this regard. Moreover, we aim to gather participant feedback to develop requirements and features that an RD repository for this use case should have.

2. RELATED WORK

RD repositories of all types, including institutional ones, with their motivations, functionalities, and overall experiences, are already well-accepted in the scholarly infrastructure. The survey from Asadi et al. (2019) includes 115 publications on the topic of IRs and lists the benefits, challenges, and motivations that universities and individual researchers seek in an IR deployment. Research dissemination, archiving of research deliverables and reputation increase of the institution and the individual research are the most cited reasons for adopting an IR in the institutional context. Furthermore, Kipnis et al. (2019) report on IR adoption trends based on the feedback from health sciences libraries, where 70% of participants already use or are in the process of deploying an IR for their needs. The survey features different aspects of such an undertaking, such as research culture (over 57% of participants do not consider an open access policy, for example), technical solutions, required repository features, etc.

MacIntyre and Jones (2016) look at another aspect of IRs: once set up, it is important to quantify the usage of the offered collections in a standard way. To this end, targeting IRs in the UK, they rely on the COUNTER ("COUNTER: Consistent, Credible, Comparable", n.d.) standard in order to provide comparable metrics available both at the repository and collection item level. This enables institutions that have IRs deployed to measure their benefits, especially in the research dissemination aspects.

When it comes to the domain requirements, the survey from Akers and Doty (2013), conducted at a research university, shows the differences in research (data) management practices among respondents from the 4 different research domains, which help identify IR services for these different research groups. On another note, one of the RDA interest groups, since 2015, is dedicated to providing recommendations to improve repository solutions ("Repository Platforms for Research Data IG", n.d.). Bringing together different repository stakeholders, this group proposed a set of use cases and matching functional requirements that RD repository

solutions should consider during development; we consulted this output during the design of our questionnaire. Moreover, Kim (2018) focuses on identifying functional requirements for RD repositories. The work considers the characteristics of RD, the available RD repository solutions, as well as the feedback from a targeted research community. The 75 requirement items that stem from this work provide a good insight about the requirements that such a repository should support, as well as a nice input to consider for RD repository-related studies, such as our own. Finally, Gordon et al. (2015), with their Databrary (IR) solution, to a similar effect, present its development as a coordinated effort between researchers and data repository developers.

As the number of available repositories increases, so does the effort required for researchers to select the ones to publish their RD on, or for publishers to setup any coordination or standardization for similar requirements. Research about quality of repositories (in general) aims to establish standards and practices to benefit research stakeholders that (want to) use these components. One of the goals of the "FAIRSharing project" is to propose a set of criteria – important to publishers and journals – that repositories can adopt in order to enable an easier navigation for researchers that want to publish their artifacts (McQuilton et al., 2020). Additionally, initiatives like CoreTrustSeal[1] or groups within Research Data Alliance[2] represent few of the initiatives that work towards adding quality-related criteria to repositories. In another context, IRs provide a nice path towards data FAIRness. In this context, considering the variety of RD and research practices across disciplines, de León and de Ferrer (2019) provide recommendations that RD repositories could adopt to become more FAIR.

IR deployment practices are already well represented, including functional, user experience (UX), technical and other aspects of the process. However, what the variety in the deployment process shows is that in such projects, taking an incremental (i.e., start small) approach, driven by the target community, is of great importance. This is the path we take with our target group – the departments of the SEE University.

3. STUDY APPROACH

Our objectives for this work are to understand RD management practices at the university level and identify corresponding (RD) services or features that an IR solution should provide. Having these goals into view, in this section we present our considerations regarding the RD activities, survey design, and survey administration.

A survey consisting of 20 questions was prepared and shared with the university teaching and research staff. As can be seen in Table 1, we mainly relied on closed-ended questions, with the few exceptions with some of the questions where the participant, by choosing the predefined option "Other", has an option to provide a free text answer. While they typically choose from a predefined list, thus dealing with a closed list of options, in a few cases, they have the option to provide an answer beyond this limitation. The questionnaire consisted of seven sections: five sections constitute the main section (see Table 1), with the remaining two used to introduce the survey topics and gather participants' email address (optional) in case they are willing to participate in a follow up for qualitative survey feedback . The first section explained the survey aim, along with the consent to participate. Next, participant demographics were gathered, such as role (teaching, research, or both), department the participant belongs to, etc. Sections 3-5 were aimed at understanding user practices in relation to data creation, preservation, and access, whereas section 6 focuses on participants' understanding of and feedback (and preferences) concerning the features that an IR should provide. Surveys with similar scope and goals always provide a useful check; we took into account the work from Guindon & Dennie (2016) and that of Hsu (2016) regarding researcher practice and repository functionalities.

---

[1] https://www.coretrustseal.org/about/
[2] https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html

Table 1. Questionnaire: Categories and questions

General Information

1	Faculty/Department
2	How long have you been working at a higher education or research institution?
3	Your job description requires

Research Data Creation

4	During your research, do you create/collect research data?
5	How often do you create/collect research data?
6	What is the format of the research data you typically create/collect?*

Research Data Preservation

7	Who manages your research data?*
8	Do you describe your created/collected research data? This could include providing title, author, publication date, subject discipline, etc.
9	How much description do you provide for your research data?*
10	Where do you store your research data?*
11	How long do you keep your data after the completion of your research or project?
12	How do you backup your research data?

Research Data Access

13	Do you share your research data? (With colleagues, publish it for a broader audience, etc.)
14	When do you typically share the research data?*
15	How do you share it?
16	Is access control important when sharing your research data?

Research Data Repository Requirements

17	Are you aware of any research data repositories in your area of research?
18	If Yes, list them
19	Is a research data repository a good support for your research activities?
20	What features would you like to see in such a repository?*

* More than one option can be applied

During the period of March - April 2019, we sent out an email invitation to the staff members of the SEE University departments to take part in the survey. We used Google Forms to administer it, which provides support for the creation, dissemination, and analysis of the questionnaire chosen as a data gathering method. The survey targeted 95 participants, across all the SEEU departments: Faculty of Contemporary Sciences and Technologies (CST), Business and Economics (BE), Languages, Cultures and Communication (LCC), Contemporary Social Sciences (CSS), and Faculty of Law. Participants were involved in teaching and/or research, including those with full- and part-time arrangements.

After the last reminder to complete the survey, the response rate reached satisfactory levels. There was a 54% response rate considering the full time members, whereas, both full- and part-time staff members generated a response rate of 45%. Table 2 shows the work experience and teaching and/or research engagement for the survey participants across the five departments.

4. FINDINGS AND DISCUSSIONS

The survey was conducted at SEE University with a total of 43 participants. The faculties of CST, BE and LCC participated equally with almost 30% each, whereas the Faculty of CSS and the faculty of Law participated cumulatively with little over 10%. Almost an absolute majority of participants, 93%, have more than 10 years of work experience at a higher education institution. From the total number, 86% are engaged in teaching and research, whereas the remaining engage only in teaching. From the received feedback, 93% of participants stated that during their research they create or collect RD. Participants typically create RD according to project needs (74%), and rarely on a monthly or weekly basis (10%).

Current research trends increasingly revolve around RD, for different motivations and activities. These trends are also present at the SEEU, to a different extent across its departments. This is an expected pattern as certain departments are more RD-driven (create or collect, analyze, persist, reuse, etc.) than others. This is the first outcome that the survey conveys.

Table 2. Survey participants' characteristics

Department	#	Higher education or research experience?		Your job description requires	
Faculty of Business and Economics	12	12	>10 years	10	Teaching and Research
				2	Teaching
Faculty of Contemporary Sciences and Technologies	11	10	>10 years	9	Teaching and Research
		1	6-10 years	1	Teaching
				1	Research
Faculty of Contemporary Social Sciences	2	2	>10 years	1	Teaching and Research
				1	Teaching
Faculty of Languages, Cultures and Communication	14	13	>10 years	12	Teaching and Research
		1	2-5 years	2	Teaching
Faculty of Law	4	3	>10 years	4	Teaching and Research
		1	6-10 years		

More diverse responses were seen in terms of the RD format that participants typically create or collect (Figure 1). As can be seen, the most popular formats of RD that participants work with are spreadsheets (47%), text documents (42%), and relational databases (40%). Taking into view the different domains represented in the survey, we encounter other formats, such as (programming) source code, comma-separated lists, or even media types.
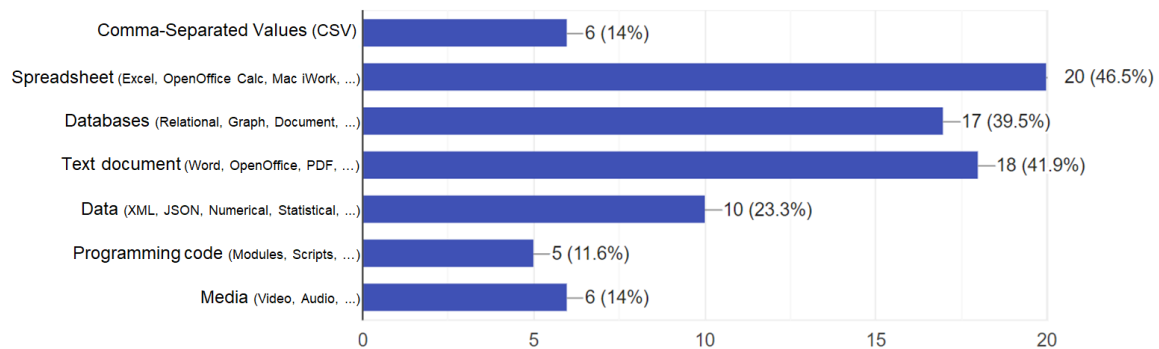

Figure 1. CSV, spreadsheets, and databases are the preferred data formats

Regarding RD preservation activities, 95% of participants stated that they manage the research data individually (Q7 from Table 1). When it comes to storage mechanisms (Q10), computer hard drives or external portable devices are the most preferred solutions (91%). In addition to this approach, 30% rely on cloud storage solutions such as Dropbox and Google Drive. Only 10% reported specialized services – data repositories – as a means for data storage. It is worth mentioning that for these questions participants had the possibility to choose more than one option, as this reflects the cases where they rely on multiple mechanisms to store their RD. We see almost the same distribution when RD backup is in question. The period of data storage is also diverse among respondents. Almost half of the respondents (51%) state that they do not know exactly how long they should keep their RD, while 30% keep it between 1 and 3 years, and 16% want to keep it for indefinite periods. Figure 2 provides all the details.

The prevalence of capable (yet affordable) computers makes them the first candidate that (almost by default) staff members turn to for data storage, including RD backup. Moreover, it seems likely that they also choose other complementary (mainly cloud-based) solutions. There is an obvious lack of repositories that could support the RD storage activities, either at individual, departmental or at a university level.
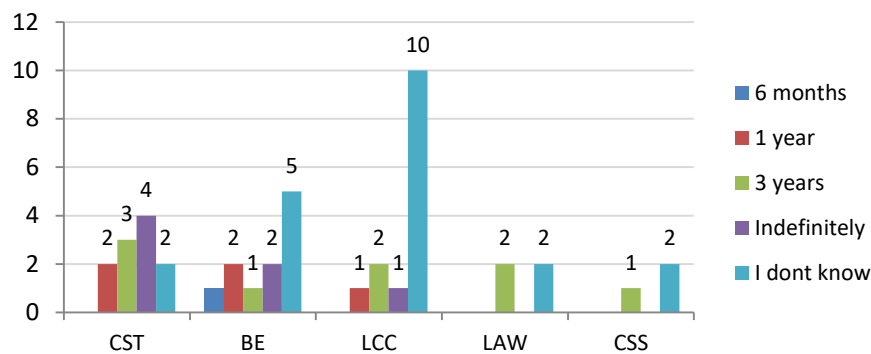
Figure 2. Time preferences for preserving RD

A good practice that we observed with the participants is that of providing accompanying metadata to RD during the different activities, such as data creation, collection or preservation. Descriptive metadata, such as title, author, subject terms, publication date, etc., seem to be well adopted with the participants (in 70% of the cases). The other metadata types we ran into is used to describe RD in the form of questionnaires or interviews (54%), those that include citation information (51%), or even metadata that describe the methodology used to create the RD (49%). In 9% of the responses we encountered the usage of metadata to describe the development environment used to create the RD. This is important to note as metadata play a valuable role during research reproducibility.

After the data creation/collection and preservation categories, we wanted to know about the RD access practices of this community, i.e., RD dissemination and access control practices when sharing or giving access to their data. A majority, 81% of the participants declared that they need to share the RD with colleagues or publish it for a broader audience. This need, however, is not present only during research/project duration. In this way, in 47% of the cases the RD is being shared after the research/project completes, whereas similar evidence (47%) is seen with data sharing triggered by request from peers or colleagues. Even though at a lesser extent, survey participants also report on RD dissemination in earlier stages of researcher, such as during (12%) or right after data creation/collection (16%).

When it comes to dissemination methods, 49% use mainly the email attachment for small datasets, 23% rely on external or portable devices, whereas 16% see cloud storage services as a method of choice. Another important issue in this context is that of RD access control. We see that 56% of survey participants prefer to give different access levels to different people, 14% prefer to give full access to everyone accessing their data, and 30% are not sure about the application of access control for their RD (see Figure 3).

The survey informs us that the RD dissemination in the largest part is supported by email attachments and external/portable storage solutions. Although the practice of storing it in more established cloud-based services is present, the lack of a departmental or institutional repository for RD is obvious.
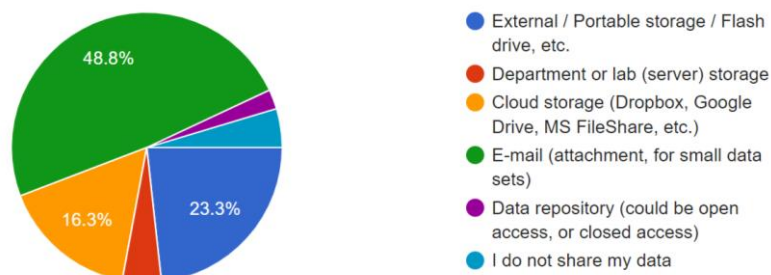

Figure 3. Data sharing practices

Given the identification of practices, as well as needs and requirements during creation/collection, preservation and dissemination of RD, we were interested to know how much the participants are aware of RD repositories. The survey results show that 79% of the participants are not familiar with the existence and features of such repositories. In any case, 47% of staff members see such a repository as a support to their

research activities, while 49% are not sure about it. Even though at a small extent (2 to 3%), we noticed that some of the participants are familiar with RD repositories for different data activities, including storage, sharing, as well as data backup. The survey results for this part show a lack of understanding for how such a repository could support one's research activities. Coupled with the current RD practices for the different departments (e.g., sharing RD via email), staff members are unsure about a potential RD repository deployment at a departmental or institutional level.

After exploring the RD practices of this community, we focused on the potential role and feature set that would be desirable for it. As shown in Figure 4, 54% of the participants are interested in data preservation features (including its metadata description), 47% of them need support during RD metadata creation and documentation, 59% are interested in features that support RD dissemination, including access control, whereas 33% see backup as an important features in such a repository. For this question (Q20), participants could select more than one option.
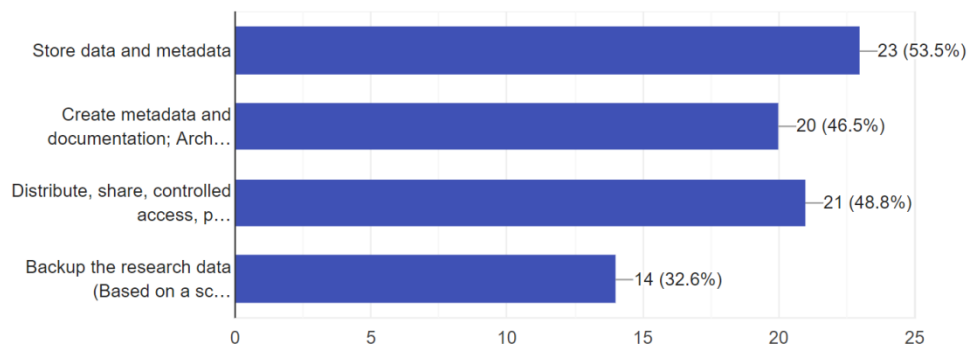

Figure 4. Features participants need in a repository

Up to this point we explored the RD practices and sought-after features for an institutional RD repository for staff members from different departments. In any case, similar requests from other stakeholders at the university, such as institutional/departmental policies, should be able to be supported by the RD repository solution. For example, while giving full access to RD could be preferred by the university, the research community might apply a "publication ban" for a certain time period on all RD created/collected of a research project. The repository of choice should be able to accommodate similar cases.

5. CONCLUSION AND FUTURE WORK

IRs represent an important component of a contemporary research infrastructure, including deployments for universities and research centers. In this paper, we focused on a university and its five departments as a use case, and aimed to identify the RD management practices for, as well as explore the role of IRs as a means of support for those practices.

Our survey targeted different research disciplines that match the different university departments. As expected, the output showed different RD practices across the departments. Although to various extent, one aspect that emerged across all departments was the increasing role and presence of RD. Moreover, the results showed a rich set of RD-related activities that the participants conduct as part of their teaching and research activities. While encouraging, it is clear that a RD IR would be a great support and efficiency boost for these (and new, yet-to-be adopted) data activities (only think at this point about such a repository to substitute sharing via email attachment or assigning different access levels to RD). This, in turn, entails RD management support, which the participants provided via the different parts of the survey, in the form of features that an IR should provide.

The limitations of this paper relate to the scope of the work, namely the target group and the RD activities covered in the survey. For the former, familiarity with the topic as well as the current role of RD across the departments impacted the results regarding some of the questions (often participants do not think about RD as an "independent", publishable research entity, thus they do not have much feedback with such parts of the survey). For the latter, we stayed on the most common (basic) RD activities, and did not explicitly include

those for data analysis, processing, archive, etc. While this was a conscious decision on our side in order to include the different departments, it left our important RD practices and repository features out. Data analysis, processing, etc., are important activities that we did not encounter (and did not prompt feedback for) with our target group. This is something we need to consider in our future work in order to have a more complete picture of research activities at this university. Moreover, our future research includes enriching the list of participants with other universities and departments in order to treat comparative aspects of RD practices and matching repository features. Finally, in addition to the RD activities covered in this survey, we plan to include analysis and processing, especially if more computationally-driven communities become part of the study.

REFERENCES

Akers, K G., & Jennifer D. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation* 8, no.2, 5-26.

Alcalá Ponce de León, M. & Anglada i de Ferrer, L. M., (2019). FAIR x FAIR. Feasible, Affordable and Implementable Requirements for a FAIR research data repository. *Report,* p. 44.

Asadi, S., Abdullah, R., Yah, Y. & Nazir, S. (2019). Understanding Institutional Repository in Higher Learning Institutions: A Systematic Literature Review and Directions for Future Research, in *IEEE Access*, vol. 7, pp. 35242-35263. COUNTER: Consistent, Credible, Comparable. (n.d.). Retrieved from https://www.projectcounter.org/.

FAIR Principles. (n.d.). Retrieved from https://www.go-fair.org/fair-principles/.

Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E., & Gilmore, R. O. (2015). Researcher-Library Collaborations: Data Repositories as a Service for Researchers. *J. of Librarianship and Scholarly Communication,* 3(2), eP1238.

Guindon, A., & Dennie, D. (2016). Concordia University Research Data Management Survey 2015-16. Available at: https://spectrum.library.concordia.ca/982722/2/RDM-Concordia-Questionnaire-Final.pdf [Accessed: 2 Oct. 2019].

H2020 Programme. (2017). Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 [online] p. 8. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [Accessed: 2 Dec. 2019].

Hsu, L. (2016). RDA: Functional Requirements for Research Data Repository Platforms. Retrieved from https://my.usgs.gov/confluence/display/cdi/RDA%3A+Functional+Requirements+for+Research+Data+Repository+Platforms.

Kim, S. (2018). Functional requirements for research data repositories. *International Journal of Knowledge Content Development & Technology* 8, no. 1, 25.

Kipnis, D. (2018). Medical Institutional Repositories in a Changing Scholarly Communications Landscape. *Against the Grain.* 30(4):33-34,36.

Lin J, & Strasser C. (2014). Recommendations for the Role of Publishers in Access to Data. *PLoS Biol* 12(10): e1001975.

Luther, J. (2018). The Evolving Institutional Repository Landscape. *ACRL/Choice publisher.*

Lynch, C. A. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age portal, *Libraries and the Academy* 3, no. 2: 327-336.

MacIntyre, R., & Hilary, J. (2016). IRUS-UK: Improving understanding of the value and impact of institutional repositories. *The Serials Librarian* 70, no. 1-4: 100-105, DOI: 10.1080/0361526X.2016.1148423.

McQuilton, P., Sansone, S.-A., Cousijn, H., Cannon, M., Chan, W. M., Carnevale, I., … Threlfall, J. (2020). *FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter*. OSF.

Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, et al. (2013). Making Research Data Repositories Visible: The re3data.org Registry. *PLoS ONE* 8(11): e78080. doi:10.1371/journal.pone.0078080.

Repository Platforms for Research Data IG. (n.d.). Retrieved from https://www.rd-alliance.org/groups/repository-platforms-research-data.html.

Hey, T. (2016). The Fourth paradigm - data-intensive scientific discovery and open science. *Book of abstracts,* Barcelona: Barcelona Supercomputing Center, p. 34-36.

Welcome to DataCite. (n.d). Retrieved from https://datacite.org/.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data 3*, 160018 doi:10.1038/sdata.2016.18.