How to Improve Research Data Management

The Case of Sciebo (Science Box)

Konstantin Wilms[1(✉)], Christian Meske[1], Stefan Stieglitz[1], Dominik Rudolph[2], and Raimund Vogl[2]

[1] Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Essen, Germany
{konstantin.wilms, christian.meske, stefan.stieglitz}@uni-due.de
[2] ZIV-Centre for Applied Information Technology, University of Muenster, Münster, Germany
{d.rudolph, r.vogl}@uni-muenster.de

Abstract. The digitalization of research processes has led to a vast amount of data. Since third-party funding institutions progressively set standards and requirements regarding the handling of such data, research data management has become important in the context of international research collaboration projects. Simultaneously, adequate collaboration systems are needed to support scientists in this context. In this paper we discuss existing standards for research data management in the context of third-party funding and how cloud technology could support the fulfillment of existing provisions.

Keywords: Cloud computing · Usability · Research data management · Technology adoption

1 Introduction

Malcom Read, executive secretary of the Joint Information Committee (JICS), stated that "We need to move away from a culture of secrecy and towards a world where researchers can benefit from sharing expertise throughout the research lifecycle" [20].

Whether in the social, behavioral, physical or computer sciences, data have always been the source of all empirical knowledge. For researchers their data are essential since they are required to prove, disprove or replicate empirical statements. For this reason, research data need to be managed professionally, in order to support efficient and effective research projects. Today, there are plenty of documented cases in which researchers lost their data or refused to disclose their research data (e.g. [32]). This is problematic, since the reproducibility of data plays a key role in many scientific fields and moreover is the only source of credibility. Although adequate data management has been an issue for a while, the responsibility for storing and disclosing data still lies with the researcher. To avoid the discussion of questionable results e.g. generated by impure data or the well-known publication bias [9], several journals now require authors to share their data sets as condition of publication [19, 27]. At the same time

S. Yamamoto (Ed.): HIMI 2016, Part II, LNCS 9735, pp. 434–442, 2016.
DOI: 10.1007/978-3-319-40397-7_41

third-party funding institutions started to set up guidelines establishing management policies for research data. Different guidelines and requirements from different funding institutions have made it difficult for the researchers to practice proper research data management (RDM). Although the number of platforms supporting the data management process is increasing, various studies indicate that there is still a lack of adoption among researchers [10]. One reason might be, that numerous departments and universities already run their own infrastructures (e.g. [11, 13, 26, 31]). While these infrastructures are generally used to provide cloud technologies, features which help researches to improve their data management are still missing.

In our paper, we focus on existing standards for RDM in the context of third-party funding and how cloud technology could support the fulfillment of existing provisions. We compare three major research funding institutions from North America (USA), Australia and Europe (Germany) in terms of requirements regarding RDM. In this first investigation we analyzed documents published by the National Science Foundation (NSF), Australian Research Council (ARC) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) with regard to requirements for research proposals for funding. Furthermore, we take the users perspective into account and focus on factors and barriers diminishing the acceptance of such systems. Additionally, we analyze, if and how sciebo ("science box"), an on-premise cloud service hosted by universities and used by over 5,000 researchers in Germany, can support scientists to meet existing RDM requirements. We especially focus on the following research questions: Which claims result from the guidelines of third-party funding institutions and from the needs of researchers for dealing with research data? How could an infrastructure like sciebo be implemented to deal with these requirements?

The remainder of the paper is structured as follows. In the next chapter the authors present some basic definitions and background knowledge about RDM and the sciebo cloud service, Following, the requirements of three different RDM standards are described and it is discussed how a system such as sciebo could support the fulfillment of the requirements. The paper ends with a conclusion.

2 Literature Review

2.1 Research Data Management

So far, there is no uniform definition of RDM on which researchers of all disciplines do agree on. One common definition describes RDM as "the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results" [33]. RDM is strongly related to the notion of "scientific data curation" which means to "collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society" [14]. According to [7] "research data" cover "any research materials resulting from primary data collection or generation, qualitative or quantitative, or derived from existing sources intended to be analyzed in the course of a research project". Data are the base of scientific communication and cover numerical data, textual data, digitized materials, images recordings or modeling scripts [7].

RDM has the potential to facilitate the entire research process and to support the efficient utilization of research data. By the disclosure of the data, the process becomes more transparent [15]. This is a basic imperative, to support the reproduction of research processes, which is a core principle in scientific research. Transparency in research helps researchers to become more resistant against the allegation of misconduct [15]. Providing access to research data has proved to be useful for scientists as sharing research data with the community may result in higher citation rates [23]. Furthermore it helps to overcome bottleneck effects, which for example could show up, if research data are only represented in the narrow context of a specialized topic [34]: "For example, a dataset collected by agronomists who are researching water quality may also be used by earth and atmospheric scientists to improve the accuracy or to validate the output of climate models" [34].

According to the Long Tail theory described by [21], shared research data have the potential to provide endless knowledge as the data are discovered and used by new audiences. Currently, there is a growing market, where several research teams migrate to basic platforms, allowing them to perform RDM and to share their data with the scientific community [1].

While the number of scientist practicing RDM is increasing [8, 17], different publications indicate, that there is still a huge mistrust when it comes to record, preserve, and share research data [5, 24, 25]. In a study of [10], the majority of researchers claimed to miss appropriate technical infrastructures for RDM, fitting all their needs. [2] found out, that a significant number of researchers need up to 100 GB of storage capacity, in order to store all their research data. Researchers also concern about ethical aspects [2]. Despite the advantages that arise from outsourcing data, public data storage services could not used without risk. Today there are still uncertainties about how data copyrights are protected in public cloud storages [12]. It is legally questionable to share or store data externally when these data are collected on the basis of the waiver of disclosure to third parties.

In addition to technical and ethical barriers there are non-technical barriers regarding a structured RDM [10]. Such non-technical barriers are for example the fear of having to compete with colleagues as well as the loss of control over the own data. Also a lack of trust to the operator of the system was reported as common reason for rejection [10]. While researchers see the added value of systematic backups and long-term storage, as they are given in cloud systems, there is still a huge number of researchers rejecting the concept of shared research data [10].

2.2 Sciebo – The Campus Cloud

The history of sciebo started in 2013 when [30] found out, that the academic community in North Rhine-Westphalia (Germany) expressed the need for an in-house cloudservice. At this time, the market was already dominated by commercial cloud services, founded by American companies like Google, Dropbox or Microsoft. Researchers expressed their concerns about privacy issues and asked for a private infrastructure placed within Germany. As a consequence the cloud infrastructure 'sciebo' was built up. Today sciebo, which is short for 'science box', is a running

infrastructure providing access to 23 academic institutions. The service is free to use and provides 30 GB storage capacity for individuals such as students as well as the academic and administrative staff of the participating institutions. In addition, project groups can apply for work boxes of up to 1 TB [28]. Up to 500,000 potential users have access to the platform. The data are stored under German data protection law. Besides the opportunity to store data, sciebo offers functions for sharing folders or set them public. Public data can also be seen and downloaded by persons who have no sciebo-account. The system, which is based on ownCloud open source software, does not offer data management functions yet. Students who finish their academic careers are given six months transfer time before their data are deleted.

3 Requirements of a Research Data Management System

3.1 Current Requirememements for Research Data Management

The following literature review takes the requirements of the three major foundation institutions into account. One of the institutions we looked at in this review is the National Science Foundation (NSF). The NSF is the largest science foundation in the United States with a promotional volume of 6.9 billion US Dollar in 2010 [29]. For Europe the 'Deutsche Forschungsgemeinschaft' (DFG) is the major funding institution in academia with a promotion budget of 2.73 billion Euro (approx. 3 billion US Dollar) in 2014 [25]. The third institution taken into account is the 'Australian Research Council' (ARC) which belongs to the major research councils in Australia [24]. Each funding institution has set up individual guidelines dealing with the topic of how RDM should be realized. In this work, we analyze the guidelines and compare the requirements. By overseeing the guidelines, twelve categories were identified (see Table 1). In the following part we show how the individual guidelines deal with the topics of data management plans (DMP), duration of storage, sharing of primary data, approaches for information collection, data standards, data security, collaboration tools for data, education in RDM, and ethics & legislation.

Table 1. Comparison of the guidelines of DFG, ARC and NSF

	Technical aspect	DFG	ARC	NSF
Data Management Plan		x	x	x
Replicability/Sharing of primary data	x	x	x	x
Duration of storage	x		x	
Approaches for information collection	x	x		
Data standards		x	x	
Data security and safety	x		x	
Collaboration tools for data	x	x		
Education in research data management		x	x	
Ethics and legislation		x	x	x

The documents comprising grant conditions and funding rules define the sharing of primary data as mandatory unless ethical or confidentiality issues prevent this. A comprehensive document labeled DMP is a required part of any proposal at the NSF including e.g. types of data, policies for access and sharing as well as plans for archiving data. Also DFG states that such a document should be included and additionally ask for e.g. data quality management, storage place, duration of access to research data, and conditions of re-use for other researchers. In contrast, the ARC considers data management planning as important but does not specify the need for a DMP. Rather, it strictly defines that research data must be retained for at least five years and should be made available for use of other researchers. The identified documents published by the DFG cover some other aspects that could not be found for the other institutions. For example, the DFG does not only require the publication of primary data, they also ask for suitable repositories and databases. Furthermore, the DFG asks as part of the proposal what implementations and techniques will be used for research data collection and processing. While the ARC emphasizes the role of researchers holding primary data including security and confidentiality aspects, the DFG maintains the education of staff in RDM but does not defer to security questions. A last example unique to the DFG is the requirement of internal collaboration tools to enable research data sharing, which is not mentioned by either ARC or NSF.

3.2 Cloud-Based Research Data Management Systems - The Case of Sciebo

This section deals with the question if and how standard cloud services among universities can support the RDM process. Since cloud services are generally based on different solutions and infrastructures, we use sciebo as an example and check if the cloud has the potential to fit the requirements pointed out in Sect. 3.1. To fit the requirements, cloud services like sciebo need to fulfill at least the technical requirements.

The first technical requirement is ‘Replicability/Sharing of primary data’. Sciebo does support two main functions which allow the user to share the data. The first function enables the users to share data among each other within sciebo. The second feature enables the user to set data public and share the data with non-sciebo-users. This option of external access is provided by sending out an http-link pointing to the data archive. Both features fit the concept of the requirements set by the three institutes. In terms of ‘duration of data’ sciebo misses the requirements. As a cloud service, sciebo provides all the technical requirements which are necessary to store long-term data. Since the service is limited to students and employees only, the users lose the right to use sciebo after e.g. graduating. Yet it is not possible to create internal relations between data stored in the system as it is required by the DFG (approaches for information collection). When it comes to ‘Data security’, as it is required by ARC, sciebo is well-positioned. The service is running under a restrictive national data protection law and uses high level security standards. Here the cloud service meets all the goals set up by the funding institute. The last technical aspect required by DFG is the need for collaboration tools. This means especially the possibility to use wikis,

blogs or data tracking tools. Sciebo currently does not support collaborative functions. However, technically there already exist some collaboration features that could be enabled in the near future.

Overall sciebo, which is representing the concept of an academic in-house cloud service, does already support some basic requirements which are necessary for being used as an RDM tool.

3.3 How to Support User Adoption

While the technical requirements are essential to fulfill the guidelines set by DFG, NSF and ARC, there are also requirements set up by the users. According to the findings of [10] the user expects that tools and services of the given RDM system are aligned to researchers discipline specific workflow. Often users require various functionalities that could be selected based on a 'cafeteria model' which allows the user to pick and choose from a set of services. Another crucial aspect is, that the researchers need to be set in a state of mind where they have the feeling of being in control over the process. They need to be awarded of 'what happens to their data, who has access to it, and under which conditions' [10]. 'Consequently, they want to be sure that whoever is dealing with their data (data centre, library, etc.) will respect their interests' [10]. To overcome the problem [2] is recommending a motivation system, where the user gets benefit by practicing RDM.

4 Discussion

RDM is a growing topic within scientific debates. According to the Horizon 2020 report the current debate is mostly focusing on the aspects of open data access and long term data storage [6]. While scientist demand improvements, it seems that universities and higher education institutes have mostly ignored the boat of this current trend. Yet RDM is mostly discussed in disciplines like medicine and microbiology [4]. However, this is a step forward to create a multirelational system among all disciplines and different workflows used by different disciplines.

[3] sees the responsibility to press ahead the implementation of such systems for the libraries. As the main competences for the librarians [3] sees the knowledge in archiving data for a long-term period and as well the competence of standardizing meta-data. Another important aspect pointed out by [3] is the implementation process. Information Systems (IS) as a discipline is therefore challenged to bring in their knowledge and competencies in order to design adequate RDM-systems. When it comes to user adoption problems or to design specific usability questions, IS researchers should develop concepts how to push up the implementation process. As a third party administrators of running academic infrastructures need to be involved in the process as well. Since the universities normally run their own in-house infrastructures, which already fulfill technical requirements partly, these could be used to support the development process. Platform or software services could be set up on the base for running infrastructure services. To foster these collaborative development processes, universities need to make investments in the future and improve their infrastructures.

Another crucial problem in RDM is the negative attitude of researchers towards an open data process. As shown in this paper, there is still a huge number of researchers which avoid to share their research data among the scientific community. It is important that the researchers join the RDM process as early as possible to gain trust in the system and overcome mentally barriers [10]. Another idea which should be considered when it comes to user adoption is to support motivational processes by incentive systems [2]. An exemplary instrument could be gamification, which has been proved to increase the activities of employees in new systems [18].

5 Conclusion

Summarizing, RDM becomes increasingly relevant for researchers. Nevertheless, the requirements differ in terms of the level of detail. Also, clear standards (e.g. for storage repositories) that can support researchers are mostly lacking. The NSF focuses mainly on the Data Management Plan (DMP), which allows (even forces) applicants to define most details themselves. The ARC does not require a formal DMP but defines a few necessities (e.g. duration of data retention). However, the ARC seems to make least requirements for research funding in terms of RDM. On the other hand, the DFG includes more aspects that are not considered by the other institutions (e.g. data management education, collaboration, as well as data collection and processing techniques) and hence, gives more weight to data management in its funded projects. Since sciebo suits several of these requirements pointed out in this research, it has the potential to provide a suitable infrastructure, through which RDM can be effectively supported.

References

1. Amorim, R.C., Castro, J.A., de Silva, J.R., Ribeiro, C.: A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential. In: Rocha, A., Correia, A.M., Costanzo, S., Reis, L.P. (eds.) New Contributions in Information Systems and Technologies. AISC, vol. 353, pp. 101–111. Springer, Heidelberg (2015)
2. Bauer, B., Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, C., Maly, N., Mühlegger, J. M., Preza, J. L., Sánchez Solís, B., Schmidt, N., Steineder, C.: Forschende und ihre Daten. Ergebnisse einer österreichweiten Befra-gung−Report 2015. Version 1.2
3. Ball, R., Wiederkehr, S. (eds.) Vernetztes Wissen. Online. Die Bibliothek als Management aufgabe: Festschrift für Wolfram Neubauer zum 65. Geburtstag. Walter de Gruyter GmbH & Co KG (2015)
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: GenBank. Nucleic Acids Res. **36**(1), D25–D30 (2008)
5. Bukavova, H.: Supporting the initiation of research collaborations. In: Jena Research Papers in Business and Economics, vol. 64 (2009)
6. Commission, E.: Guidelines on open access to scientific publications and research data in horizon 2020. Technical report (2013)

7. Corti, L., Van den Eynden, V., Bishop, L., Woollard, M.: Managing and Sharing Research Data: A Guide to Good Practice. Sage Publications Ltd, New York (2014)
8. da Silva, J.R., Barbosa, J.P., Gouveia, M., Ribeiro, C., Lopes, J.C.: UPBox and DataNotes: a collaborative data management environment for the long tail of research data. In: iPRESS2013: Proceedings of the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal (2013)
9. Easterbrook, P.J., et al.: Publication bias in clinical research. Lancet **337**(8746), 867–872 (1991)
10. Feijen, M.: What Researchers Want. SURF-foundation, Utrecht (2011)
11. Hager, R., Hildmann, T., Bittner, P.: Ein Jahr mit ownCloud – von der Planung bis zur Neustrukturierung. In: Kao, O., Hildmann, T. (eds.) Cloudspeicher im Hochschuleinsatz: Proceedings der Tagung "Cloudspeicher im Hochschuleinsatz" am 05. und 06. Mai 2014 am IT-Service-Center (tubIT) der Technischen Universität Berlin, Universitätsverlag der TU Berlin (2014)
12. Hilber, M., Reintzsch, D.: Cloud Computing und Open Source-Wie groß ist die Gefahr des Copyleft bei SaaS? Comput. Und Recht: Forum für die Praxis des Rechts der Datenverargeitung, Inf. und Autom. **30**(11), 697–702 (2014)
13. Hildmann, T., Kao, O.: Deploying and extending on-premise cloud storage based on ownCloud. In: 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops. IEEE (2014)
14. Johns Hopkins University. http://www.dataconservancy.org/home. Accessed 4 Mar 2009
15. Joshi, M., Krag, S.S.: Issues in data management. In: Spier, R.E., Bird, S.J. (eds.) Science and Engineering Ethics, vol 16:4, pp. 743–748. Springer, Heidelberg (2010)
16. Library Trends, vol. 57:2, pp. 191−201. Johns Hopkins University Press, Illinois (2008)
17. Lyon, L.: Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report (2007)
18. Meske, C., Brockmann, T., Wilms, K., Stieglitz, S.: Gamify employee collaboration – a critical review of gamification elements in social software. In: ACIS (2015)
19. Nature Journal. Authors & Referees, Editorial Policies, Availability of data & materials. http://www.nature.com/authors/editorial_policies/availability.html. Accessed 4 Mar 2009
20. 'New advice for universities in light of the Climate Change Emails Review' JISC news release July 2010. http://www.jisc.ac.uk/news/stories/2010/07/opendata.aspx
21. Palmer, C.L., Cragin, M.H., Heidorn, P.B., Smith, L.C.: Data curation for the long tail of science: the case of environmental sciences. In: Third International Digital Curation Conference, Washington, DC (2007)
22. Peterson, M., Zasman, G., Mojica, P., Porter, J.: 100 year archive requirements survey. In: SNIA's Data Management Forum, p.1. SNIA (2007)
23. Piwowar, H.A., Day, R.S., Fridsma, D.B.: Sharing detailed research data is associated with increased citation rate. PLoS ONE **2**(3), e308 (2007)
24. Savage, C.J., Vickers, A.J.: Empirical study of data sharing by authors publishing in PloS journals. PLoS ONE **4**(9), e7078 (2009)
25. Sayogo, D.S., Pardo, T.A.: Exploring the determinants of scientific data sharing: understanding the motivation to publish research data. Gov. Inf. Q. **30**, 19–31 (2013)
26. Schlitter, N., Yasnogorbw, A.: Sync&Share: a cloud solution for academia in the state of Baden-Württemberg. In: Kao, O., Hildmann, T. (eds.) Cloudspeicher im Hochschuleinsatz: Proceedings der Tagung "Cloudspeicher im Hochschuleinsatz" am 05. und 06. Mai 2014 am IT-Service-Center (tubIT) der Technischen Universität Berlin, Universitätsverlag der TU Berlin (2014)
27. Science Magazine. General Information for Authors. Conditions of Acceptance. http://www.sciencemag.org/about/authors/prep/gen_info.dtl#datadep. Accessed 4 Mar 2009

28. Stieglitz, S., Meske, C., Vogl, R., Rudolph, D.: Demand for cloud services as an infrastructure in higher education. In: Icis (2014)
29. Suresh, S.: Biography. http://www.nsf.gov/news/speeches/suresh/suresh_bio.jsp Accessed 10 Jan 2016
30. Vogl, R., Angenent, H., Bockholt, R., Rudolph, D., Stieglitz, S., Meske, C.: Designing a large scale cooperative sync&share cloud storage plattform for the academic community in Northrhine-Westfalia. In: Proceedings of EUNIS 2013 Congress 1(1), (2013)
31. Vogl, R., Rudolph, D., Thoring, A., Angenent, H., Stieglitz, S., Meske, C.: How to build a cloud storage service for half a million users in higher education: challenges met and solutions found. In: HICCS 2016 Proceedings of the 49th Hawaii International Conference on System Sciences, pp. 5328−5337 (2016)
32. Voorbrood, C.M. Voer voor psychologen. Archivering, beschikbaarstelling en hergebruik van onderzoeksdata in de psychologie, Aksant Academic Publishers (2010)
33. Whyte, A., Tedd, J.: Making the case for research data management. Digitar curation centre, edinburgh (2011)
34. Witt, M.: Institutional repositories and research data curation in a distributed environment. In: Library Trends, vol 57:2, pp. 191–201. John Hopkins University Press and the Graduate School of Library and Information Science, Illinois (2008)