A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential

Ricardo Carvalho Amorim[1], João Aguiar Castro[1], João Rocha da Silva[1], and Cristina Ribeiro[2]

[1] Faculdade de Engenharia da Universidade do Porto/INESC TEC
{ricardo.amorim3,joaoaguiarcastro,joaorosilva}@gmail.com
[2] DEI—Faculdade de Engenharia da Universidade do Porto/INESC TEC
mcr@fe.up.pt

Abstract. Research data management is acknowledged as an important concern for institutions and several platforms to support data deposits have emerged. In this paper we start by overviewing the current practices in the data management workflow and identifying the stakeholders in this process. We then compare four recently proposed data repository platforms—DSpace, CKAN, Zenodo and Figshare—considering their architecture, support for metadata, API completeness, as well as their search mechanisms and community acceptance. To evaluate these features, we take into consideration the identified stakeholders' requirements. In the end, we argue that, depending on local requirements, different data repositories can meet some of the stakeholders requirements. Nevertheless, there is still room for improvements, mainly regarding the compatibility with the description of data from different research domains, to further improve data reuse.

1 Introduction

The number of published scholarly papers is steadily increasing, and there is a growing awareness of the importance, diversity and complexity of data generated in research contexts. The management of these valuable assets is currently a concern for both researchers and science managers who have to streamline scholarly communication, while keeping record of research contributions and ensuring the correct licensing of their contents [1,2]. A gradual increase in the number of research publications, combined with a strong drive towards open access policies [3,4], have led to the development of several open-source platforms for managing bibliographic records. At the same time, academic institutions have new mandates for the management of their research data [5], increasing the demand for infrastructures that can support these activities.

When faced with the many alternatives currently available, it can be difficult for institutions to choose a suitable platform to meet their specific requirements. Thus, several comparative studies between existing solutions were carried out in order to evaluate different aspects of each implementation [6,7,8].

Á. Rocha et al. (eds.), *New Contributions in Information Systems and Technologies*, Advances in Intelligent Systems and Computing 353, DOI: 10.1007/978-3-319-16486-1_10

It has been shown that research publications that provide access to their base data yield consistently higher citation rates than those that do not [9]. As data management becomes an increasingly important part of the research workflow [10], solutions designed for managing research data are being actively developed by open-source communities. As with publication repositories, many of their design and development challenges are aimed at the description and long-time preservation of research data. When comparing publications and datasets, however, different research domains have different requirements, mainly due to the structural diversity of their datasets. Metadata requirements can also vary greatly from domain to domain, requiring repository data models to be flexible enough to adequately represent these records [11].

In this paper, we present an overview of several prominent research data management platforms that can implement a part of the research data management workflow. We consider four well-known open-source platforms and evaluate them according to a set of key aspects: architecture, metadata handling capabilities, interoperability, content dissemination, search features and community acceptance. Interoperability, in particular, plays an important role in ensuring that data can be found and reused by researchers and institutions outside of the original research group. Repository directories such as *re3data*[1] and *OpenDOAR*[2] take advantage of established interoperability protocols to help researchers find relevant data repositories [12]. To account for the different requirements of common stakeholders in a data management workflow—researchers, institutions, curators, harvesters, and developers—we analyze each of the platforms according to their most common use cases. This evaluation is a result of a preliminary study of the available repository solutions, considering aspects and repositories relevant to our ongoing work, and that were not addressed in other evaluations.

The paper is organized as follows: Section 2 briefly describes the current data management practices, considering researchers without a supporting data management infrastructure, and identifies key stakeholders in this workflow that can directly influence how data is described and preserved. Section 3 shows an evaluation of four well-known platforms that are currently used to manage, and in some cases preserve, research data, taking into consideration the previously identified stakeholders' requirements. Section 4 elaborates on the use of the repository platforms in a research management workflow. Finally, Section 5 summarizes the conclusions that can be drawn from this comparison, pointing out the main advantages of each of the evaluated platforms, according to the stakeholders.

2 Research Data Management

Managing research data is nowadays a common practice to ensure that the research outputs are recorded and preserved over time [10,13]. Currently, several research teams resort to known platforms to manage their data and ultimately share it with the research community, while also guaranteeing their preservation.

[1] `http://www.re3data.org/`
[2] `http://www.opendoar.org/`

Due to their growing popularity, these platforms also tend to implement features to describe the deposited datasets, in a similar way to the established practice concerning research papers.

Several stakeholders are involved in dataset description throughout the data management workflow while taking part in their management and dissemination [10,14]. These stakeholders—*researchers*, *research institutions*, *curators*, *harvesters*, and *developers*—play a governing role in defining the main requirements of a data repository for the management of research outputs. As a key metadata provider, *researchers* are responsible for the description of their research data through domain-level records, but are not necessarily knowledgeable in data management practices. Domain-specific descriptions complement generic metadata, thus providing the much needed data production context that makes other researchers comfortable with their reuse [14]. *Institutions* are also motivated towards having their researchers' data recognized and preserved according to the requirements of funding institutions [15,16]. In this regard, they value the compliance with metadata standards to make them ready for inclusion in networked environments and therefore increase their visibility. To make sure that this context is correctly passed along with the data to the preservation stage, *curators* are mainly interested in maintaining data quality and integrity over time. They are often not experts in the research domains of the datasets that they curate, so they must work in close collaboration with researchers to produce detailed and compliant metadata records [17]. To promote long-term preservation, they are often information specialists (such as librarians and archivists) or even members of the research team with some expertise in data description. Considering data dissemination and reuse, *harvesters* can be either persons looking for specific data or services which index the content of several repositories. These services can make particularly good use of established protocols, such as the OAI-PMH [18], to retrieve metadata from different sources and create an interface to expose the indexed resources. Finally, contributing to the improvement and expansion of these repositories over time, *developers* are concerned about the underlying technologies, but mostly in having extensive APIs to promote integration with other tools.

Considering the identified stakeholders, choosing a suitable data repository is not an easy task. Thus, a summary of the overall features available in each platform can be a good starting point, as it can help institutions make an informed choice.

3 Platform Evaluation

For this comparison, we have selected four open-source data management platforms with instances running at research and government institutions, namely DSpace, CKAN, Zenodo and Figshare. Depending on the involved stakeholders, however, other repositories may also be considered in the selection of a data management platform. From a long-term preservation point of view, for example, other alternatives such as RODA [19] may be considered, since they implement

comprehensive preservation guidelines, not only for the digital objects themselves but also for their whole life cycle and associated processes. On one hand, these platforms have a strong concern with long-term preservation by strictly following existing preservation standards such as OAIS, PREMIS or METS. On the other hand, such solutions may be harder to install and maintain by institutions in the so-called *long-tail* of research data—institutions that create large numbers of small datasets, but that do not possess the necessary financial resources and preservation expertise to support a complete preservation workflow [2].

An overview of the previously identified stakeholders led us to select four main categories in which these repositories can be assessed: their architecture and implementation features, the comprehensiveness of the metadata records that they support, their data dissemination capabilities as well as their adoption from the research community point of view. Table 1 has an overview of the results of our evaluation.

3.1 Architecture

Regarding the architecture of each platform, several important aspects must be considered. From the institution's point of view, a quick and simple deployment of the selected platform can be an important aspect. Two scenarios arise in this regard: either the institution signs for an external service or installs and customizes its own repository, and thus has to support the infrastructure maintenance costs. Contracting a service provided by a dedicated company such as Figshare or Zenodo delegates system maintenance to the company for a monthly fee. The service-based approach may not be viable in some scenarios, as some researchers or institutions may be reluctant to deposit their data in a platform outside their control [20]. In this matter, both DSpace and CKAN may offer a better control over the stored data as they can be installed and run completely under the control of the research institution. As open-source solutions, they also have several supporters[3] that contribute to their expansion by developing additional plugins or extensions to meet specific requirements. DSpace, CKAN and Zenodo allow a certain degree of customization to better satisfy the needs of their users: while Zenodo allows parametrization settings such as community-level policies, CKAN and DSpace—as open source solutions—can be further customized, with improvements ranging from small interface changes to the development of new data visualization plugins [21].

A collaborative environment where teams and groups can manage the deposited resources is becoming increasingly important in the research workflows of many institutions. In this regard, both CKAN and Zenodo provide collaborative tools and allow users to fully manage their group members and policies. As some researchers may want to control the data release dates, DSpace and Zenodo also allow specifying embargo periods after which data is made available to the community.

[3] `http://ckan.org/instances/`
`http://registry.duraspace.org/registry/dspace`

Table 1. Results of the evaluation of the selected platforms

Class	Feature	DSpace	CKAN	Figshare	Zenodo
Architecture	Deployment	Installation package	Installation package	Service	Service
	Storage location	Local or remote	Local or remote	Remote	Remote
	Maintenance costs	Infrastructure management	Infrastructure management	Monthly fee	Monthly fee
	Open Source	✓	✓	✗	✗
	Platform customization	✓	✓	✗	Community policies
	Embargo period	✓	Private storage	Private storage	✓
	Content versioning	✗	✓	✗	✗
	Pre-reserving DOI	✓	✗	✓	✓
Metadata	Required fields	Title, Date of issue	Title	Author, title, categories, description	Type, DOI, author, title, description
	Exporting schemas	Any pre-loaded schema	✗	DC	DC, MARCXML
	Schema flexibility	Flexible	Flexible	Fixed	Fixed
	Validation	✓	✗	✗	✓
	Versioning	✗	✓	✗	✗
Dissemination	API	✓	✓	✓	✓
	OAI-PMH Compliance	✓	✗	✓	✓
	Faceted search	✓	✓	✓	✓
	Metadata included	✓	✓	✓	✓

3.2 Metadata: A Key for Preservation

Research data can benefit from domain-level metadata to contextualize their production [22]. While the evaluated platforms require descriptions of varying

detail when depositing datasets, most of them lack the support for domain-specific metadata schemas. In this regard DSpace is a notable exception, with its ability to use multiple schemas that can be set up by a system administrator. Both Zenodo and Figshare are able to export records that comply with established metadata schemas (Dublin Core and MARC-XML and Dublin Core, respectively), but DSpace goes further by exporting DIPs (Dissemination Information Package) that include METS metadata records, thus enabling the ingestion of these packages into a long-term preservation workflow. Although CKAN metadata records do not follow any standard schema, the platform allows the inclusion of a dictionary of key-value pairs that can be used to record domain-specific metadata as a complement to generic metadata descriptions. Neither platform natively supports collaborative validation stages where curators and researchers enforce the correct data and metadata structure, but Zenodo allows the users to create a highly curated area within *communities*, as highlighted in the "validation" feature in Table 1. If the policy of a particular *community* specifies manual validation, every deposit will have to be validated by the community *curator*. Tracking content changes is also an important issue in data management, as datasets are often versioned and dynamic. CKAN provides an auditing trail of each deposited dataset by showing all changes made to it since its deposit.

3.3 Interoperability and Dissemination

Exposing repository contents to other research platforms can improve both data visibility and reuse [10]. All of the evaluated platforms allow the development of external clients and tools as they already provide their own APIs for exposing metadata records to the outside community, but there are some differences regarding standards compliance. Only Zenodo and DSpace natively comply with the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol [18]. This is a widely-used protocol that promotes interoperability between repositories while also streamlining data dissemination, and is a valuable resource for harvesters to index the contents of the repository [23,24].

It is interesting to evaluate the ease of discovery by machine but also how easily humans can find a dataset. All three platforms possess free-text search capabilities, indexing the metadata in dataset records for use in retrieval. All analyzed platforms provide an "advanced" search feature that is in practice a faceted search. Depending on the platform, it allows users to restrict the results to smaller sets, for instance, from the Engineering domain. This search feature makes it easier for researchers to find the datasets that are from relevant domains and that belong to specific collections or similar dataset categories (the concept varies between platforms as they have different organizational structures).

3.4 Platform Adoption

As most recent platforms, all the repositories depend on numerous developers to maintain and improve their features. Looking for successful case studies, it

Table 2. Key advantages of the evaluated repository platforms

Platform	Key advantages
Figshare	– Gives credit to authors through citations and references – Can export reference to Mendeley, DataCite, RefWorks, EndNote, NLM and ReferenceManager – Records statistics related to citations and shares – Does not require any maintenance
Zenodo	– Allows creating communities to validate submissions – Supports Dublin Core, MARC and MARCXML for metadata exporting – Can export references to BibTeX, DataCite, DC, EndNote, NLM, RefWorks – Complies with OAI-PMH for data dissemination – Does not require any maintenance – Includes metadata records in the searchable fields
CKAN	– Is open-source and widely supported by the developer community – Features extensive and comprehensive documentation – Allows deep customization of its features – Can be fully under institutions control – Supports unrestricted (non standards-compliant) metadata – Has faceted search with fuzzy-matching – Records datasets change logs and versioning information
DSpace	– Can comply with domain-level metadata schemas – Is open-source and has a wide supporting community – Has an extensive, community maintained documentation – Can be fully under institutions control – Structured metadata representation – Compliant with OAI-PMH

is important to assess their impact and comprehensiveness. CKAN has several success cases with government data which are made available to the community, but misses other scenarios related to the management and disclosure of research data. The other platforms—Figshare, Zenodo and DSpace—have research data as their focus. In active use since 2002, DSpace is well known among institutions and researchers for its capabilities to deal with research publications and, more recently, to handle research data in a similar way. Considering this, DSpace may have better acceptance due to the existence of instances already in place that can ease its management and upgrade. Zenodo is a solution for the long tail of science supported by CERN laboratories, and is regarded as an environment to bring research outputs to a proper digital archive for preservation. It is therefore a strong use case, with all the currently active researchers making use of its features, regardless of their research field.

4 Ongoing Research

Most of the analyzed solutions tend to focus on holding and managing research data outputs after the results of their analysis are published. As a consequence, we highlight a lack of support for earlier stages of the researchers' activities. Capturing data and metadata at an earlier moment in the research workflow—ideally as it is produced—presents an opportunity for increasing the number of datasets that actually make it into a long-term preservation environment. Improvements in metadata quality can also be derived from an earlier deposit timing and from a centralization of the metadata records. Combined with a collaborative metadata production process carried out by researchers, such a change in practice might improve the sharing of dataset production contexts among the members of the research group and their partners. Researchers can then become more aware of the short-term advantages of building structured metadata records, instead of viewing metadata production as a process that is either mandatory or motivated by uncertain, long-term rewards—such as the possibility of others citing their datasets.

An ongoing project targeted at improving the overall availability and quality of research data—Dendro—aims at closely involving researchers in the management and description of their data, focusing on metadata recording at the early stages of the research workflow [25,26]. Dendro consists of a fully open-source environment (solution and dependencies) that combines an easy to use file manager (similar to *Dropbox*[4]) with the collaborative capabilities of a *semantic wiki* for the production of semantic metadata records. The solution aims at the description of datasets from different research domains through an extensible, triple store-based data model [11]. Curators can expand the platform's data model by loading ontologies that specify domain-specific or generic metadata descriptors that can then be used by researchers in their descriptions. These ontologies can be designed using tools such as Protégé[5], allowing curators with no programming background to extend the platform's data model. Dendro is designed primarily as a *staging environment* for dataset description. Ideally, as research publications are written, base datasets (already described at this point) are packaged and sent to a research data repository, where they go through the usual deposit workflows. In the end, the process is fast enough to enable researchers to cite the datasets in the publication itself as supporting data.

Dendro focuses on interoperability in order to make the deposit process as easy as possible for researchers. It integrates with all the repository platforms surveyed in this paper, while its extensive API makes it easy to integrate with external systems. LabTablet, an electronic laboratory notebook designed to help researchers gather metadata in experimental contexts is an example of a successful integration scenario. It allows researchers to produce metadata records using the mobile device's onboard sensors, which are then represented using established metadata schemas (e.g. Dublin Core) and uploaded to a Dendro instance for collaborative editing [27].

[4] `https://www.dropbox.com/`

[5] Available at `http://protege.stanford.edu/`

5 Conclusion

Our evaluation showed that it can be hard to select a platform over the others without first performing a careful study of the requirements of all stakeholders. The main positive aspects of the platforms we have considered are summarized in Table 2. We highlight both CKAN and DSpace's open-source licenses that allow them to be updated and customized, while keeping the core functionalities intact. Although CKAN is mainly used by governmental institutions to disclose their data, its features and the extensive API make it also possible to use this repository to manage research data, making use of its key-value dictionary to store any domain-level descriptors—but we would like to highlight that this representation does not strictly enforce a metadata schema. Curators may favor DSpace though, since it enables system administrators to parametrize additional metadata schemas that can be used to describe resources. These will in turn be used to capture richer domain-specific features that may prove to be essential for data reuse. Researchers need to comply with funding agency requirements, so they may favor easy deposit combined with easy data citation. Zenodo and Figshare provide ways to reserve a permanent link and a DOI, even if the actual dataset is under embargo at the time of first citation. This will require a direct contact between the data creator and the potential re-user before access can be provided. Both these platforms are aimed at the direct involvement of researchers in their data publication, as they streamline the upload and description processes, but they do not provide support for domain-specific metadata descriptors. A very important factor to consider is also the control over where the data is stored. Some research institutions may want to implement a solution where data is stored in servers completely under their control and to directly manage their research assets. In this sense, we highlight platforms such as DSpace and CKAN, that can be installed in an institutional server instead of relying on external storage provided by contracted services.

The evaluation of research data repositories can take into account other features besides those considered here, namely their acceptance within research communities or specific domains and their usability which have not been considered in this analysis. We have focused on repositories as final locations for research data to be deposited and not a replacement for the tools that researchers already use to manage their data—such as file sharing environments or more complex e-science platforms. Instead, we consider that these solutions should be compared to other collaborative solutions such as Dendro, a research data management solution currently under development. In this regard, we argue that flexible, customizable solutions such as Dendro can meet the institutions' needs in terms of staging, temporary platforms to help with research data management and description. This, while taking into consideration available metadata standards that can contribute to overall better conditions for long-term preservation [26].

Finally, considering small institutions that somehow struggle to contract a dedicated service for this purpose, having a wide community supporting the development of stand-alone platforms can be a valuable asset. In this regard,

CKAN may have an advantage over the remaining alternatives, as several governmental institutions are already converging to this platform for their data publishing needs.

Acknowledgements. This work is supported by project NORTE-07-0124-FEDER000059, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). João Rocha da Silva is also supported by research grant SFRH/BD/77092/2011, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

1. Lynch, C.A.: Institutional repositories: essential infrastructure for scholarship in the digital age. Portal: Libraries and the Academy (2003)
2. Heidorn, P.B.: Shedding light on the dark data in the long tail of science. Library Trends 57(2), 280–299 (2008)
3. Burns, C.S., Lana, A., Budd, J.: Institutional repositories: exploration of costs and value. D-Lib Magazine 19 (2013)
4. Coles, S.J., Frey, J.G., Bird, C.L.: First steps towards semantic descriptions of electronic laboratory notebook records. Journal of Cheminformatics 2013, 1–10 (2013)
5. Commission, E.: Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Technical Report (December 2013)
6. Fay, E.: Repository software comparison: building digital library infrastructure at LSE. Ariadne (2009), 1–11 (2010)
7. Armbruster, C., Romary, L.: Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in. International Journal of Digital Library Systems (2009)
8. Bankier, J.G.: Institutional Repository Software Comparison. UNESCO Communication and Information 33 (2014)
9. Piwowar, H.A., Day, R.B., Fridsma, D.S.: Sharing detailed research data is associated with increased citation rate. PLOS ONE 2(3) (2007)
10. Lyon, L.: Dealing with Data: Roles, Rights, Responsibilities and Relationships. Technical report, UKOLN, University of Bath (2007)
11. Silva, J.R.d., Ribeiro, C., Lopes, J.C.: Ontology-based multi-domain metadata for research data management using triple stores. In: Proceedings of the 18th International Database Engineering & Applications Symposium (to be published 2014)
12. Pampel, H., Vierkant, P., Scholze, F.: Making research data repositories visible: the re3data. org registry. PLOS ONE 8(11), 1–18 (2013)
13. Ribeiro, C., Barbosa, J., Gouveia, M., Lopes, J., Silva, J.R.D.: UPBox and DataNotes: a collaborative data management environment for the long tail of research data. In: iPres 2013 Conference Proceedings (2013)
14. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6) (2012)

15. Green, A., Macdonald, S., Rice, R.: Policy-making for Research Data in Repositories: A Guide. DISC-UK, Edinburgh (May 2009)
16. Foundation, N.S.: Grants.gov Application Guide A Guide for Preparation and Submission of NSF Applications via Grants.gov. (2011)
17. Swan, A., Brown, S.: The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Report to the JISC (2008)
18. Lagoze, C., Sompel, H.V.D., Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting. In: Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2001 (2001)
19. Ramalho, J.C., Ferreira, M., Faria, L., Castro, R.: RODA and CRiB a service-oriented digital repository. In: Proceedings of the 5th International Conference on Preservation of Digital Objects, iPRES 2008 (2008)
20. Corti, L., Eynden, V.d., Bishop, L., Woollard, M.: Managing and Sharing Research Data: A Guide to Good Practice. SAGE Publications (2014)
21. Silva, J.R.d., Ribeiro, C., Correia Lopes, J.: Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. In: iPres 2012 Conference, pp. 105–108 (2012)
22. Willis, C., Greenberg, J., White, H.: Analysis and Synthesis of Metadata Goals. Journal of the Association for Information Science and Technology 63(8), 1505–1520 (2012), doi:10.1002/asi
23. Breu, F.X., Guggenbichler, S., Wollmann, J.C.: Research and Advanced Technology for Digital Libraries. Vasa (2008)
24. Devarakonda, R., Palanisamy, G.: Data sharing and retrieval using OAI-PMH. Earth Science Informatics 4(1), 1–5 (2011)
25. Silva, J.R.: Dendro: collaborative research data management built on linked open data. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC 2014, vol. 8798, pp. 3–13. Springer, Heidelberg (2014)
26. Silva, J.R.d., Ribeiro, C., Lopes, J.C.: The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In: iPres 2014 Conference Proceedings (2014)
27. Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C.: LabTablet: Semantic metadata collection on a multi-domain laboratory notebook. In: Closs, S., Studer, R., Garoufallou, E., Sicilia, M.-A. (eds.) MTSR 2014. Communications in Computer and Information Science, vol. 478, pp. 193–205. Springer, Heidelberg (2014)