Solutions for data sharing and storage: a comparative analysis of data repositories

Joana Rodrigues[1,2][0000−0002−1309−2122] and Carla Teixeira Lopes[1,2][0000−0002−4202−791X]

[1] INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[2] Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
joanasousarodrigues.14@gmail.com, ctl@fe.up.pt

Abstract. Research data management is an essential process in scientific research activities. It includes monitoring data from the moment it is created until it is deposited in a repository so that later it can be accessed and reused by others. Sharing and reuse are the last steps in this process. It is essential to ensure that the data stored in digital repositories is well preserved in the long term and that its adequate interpretation and future reuse is guaranteed. Following this debate, questions arise related to the interoperability of systems and the suitability of platforms. In this study, we study how data management platforms can solve the problems associated with description, preservation, and access in digital media, making their usefulness evident. We identify some of the most relevant repository platforms in the scope of research data management, offering the scientific community an aggregating view of the various solutions and their main characteristics, thus aiming at a better understanding of them for their appropriate choice.

Keywords: Research Data Management, Data Repositories, Sharing

1 Contextualization

Currently, the number of articles and datasets produced in science is increasing [1]. Allied to this, we see a growing awareness of the importance of producing research data. This fact motivated the publication and sharing of data to gain a greater importance for researchers who want to see their data properly organized, stored, and described, in order to promote their sharing, reuse, and citation. The management of these resources has become a concern for researchers and research institutions. A proper data management contributes to the reproducibility of science, the reduction of duplicate efforts in data production, and the possibility of comparing results with those of peers [4,6].

Research data management can be seen as a set of policies and activities, which accompany the entire life cycle of data and which aim to ensure that they fulfill their role in the context of the research activity and, in particular, that they are preserved [1]. One of the main results of this practice is that researchers use platforms to manage their data, share them with the entire research community, and contribute to their preservation over time [1,3]. The diversity of platforms

Rodrigues, J., Lopes, C.T. (2022). Solutions for Data Sharing and Storage: A Comparative Analysis of Data Repositories. In: , et al. Linking Theory and Practice of Digital Libraries. TPDL 2022. Lecture Notes in Computer Science, vol 13541. Springer, Cham. https://doi.org/10.1007/978-3-031-16802-4_55

available makes it difficult to choose, as it becomes more complicated to choose the one that specifically meets identified needs. Some of the obstacles are in defining metadata for description and long-term preservation [1].

The abrupt growth of data production is one of the reasons that promote repository platforms to implement functionalities to describe the datasets that are deposited on them [2]. The various stakeholders in this process, whether research institutions, individual researchers, or curators, contribute to the description of the data produced. In this context, the role of curators is intrinsically linked to maintaining the accessibility, quality, and integrity of data over time [1]. Thus, the description that researchers make of their data, combined with the metadata of the datasets themselves, creates the necessary conditions for their future reuse and citation by other researchers. The support of protocols such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) for retrieving metadata from different sources and creating an interface for displaying indexed resources makes the dissemination of data more effective [5].

2 Repository analysis

The selection of platforms took into account their relevance in the context in question, recognizing their influence and usefulness in the context of research data management. Based on the Registry of Open Access Repositories (ROAR)[3] statistics, eight software solutions were selected, but this was not the only choice criterion. Priority was given to platforms that support interoperability features and also support repositories of research and government institutions. Issues associated with the description of datasets or the definition of metadata were also taken into account. Thus, the following platforms were considered: CKAN, Dataverse, DSpace, ePrints, EUDAT, Figshare, Zenodo, and Islandora.

We analyze each of the platforms according to six analysis criteria. These criteria were derived from the documentation of the platforms and the experimentation of demo instances. They are divided into six categories whose structure is inspired by the Open Archival Information System (OAIS) model [7] and can be seen in the next subsections.

2.1 Infrastructure

Basic attributes that support the functioning of the platforms.

Table 1: Comparison of platforms for the infrastructure component

	Infrastructure			
	Open Source	**Storage**	**Installation**	**Payment system**
CKAN	√	Local or remote	Installation	x
Dataverse	√	Local or remote	Intallation or service	x
DSpace	√	Local or remote	Installation or service	x
ePrints	√	Local or remote	Installation or service	x
EUDAT	√	Remote	Service	x
Figshare	x	Remote	Service	x
Zenodo	√	Remote	Service	x
Islandora	√	Local or remote	Installation	x

[3] http://roar.eprints.org/view/software/

2.2 Ingestion

Integration and interoperability within the platform itself or with other systems.

Table 2: Comparison of platforms for the ingestion component

	Ingestion		
	Interoperability	**API of deposit of data**	**Automatic import of metadata from files**
CKAN	OAI-PMH	SWORD	x
Dataverse	OAI-PMH	SWORD	√
DSpace	OAI-PMH	SWORD	√
ePrints	OAI-PMH	x	√
EUDAT	OAI-PMH	SWORD	√
Figshare	OAI-PMH	SWORD	√
Zenodo	OAI-PMH, REST	SWORD	√
Islandora	OAI-PMH	SWORD	√

2.3 Content organization and control

Structuring of content and its control within each platform.

Table 3: Comparison for the content organization and control component

	Content organization and control					
	Organization of contents	**Access granularity**	**User profile**	**Authentication**	**Support of attached documents**	**Embargo**
CKAN	Linear	Grups	√	√	√	√
Dataverse	Hierarchical	File, User	√	√	√	√
DSpace	Hierarchical	Groups	√	√	√	√
ePrints	Hierarchical and linear	User	√	√	√	√
EUDAT	Linear	User	√	√	√	√
Figshare	Linear	Institution, Publisher, Researcher	√	√	-	√
Zenodo	Hierarchical	User	√	√	√	√
Islandora	Hierarchical	User	√	√	√	√

	Content organization and control					
	Volume and size	**Eligibility of depositor**	**Language**	**Data maturity**	**Deposit elimination**	**Licensing**
CKAN	-	All allowed	All allowed	Any state	√	√
Dataverse	database:3GB, files:3GB, records:1GB	All allowed	All allowed	Any state	√	√
DSpace	-	All allowed	All allowed	Any state	√	√
ePrints	-	All allowed	All allowed	Any state	√	√
EUDAT	max 20GBper record, max 10 GB per file	All allowed	All allowed	Any state	√	√
Figshare	max 20GB per record	All allowed	All allowed	Any state	√	√
Zenodo	max 50GB per record	All allowed	All allowed	Any state	√	√
Islandora	-	All allowed	All allowed	Any state	√	√

2.4 Metadata

Data description and the ways in which it takes place on different platforms.

Table 4: Comparison for the metadata component

	Metadata			
	Schema/Standard/Model Flexibility	**Schema Export**	**Validation**	**License registration**
CKAN	√	x	x	√
Dataverse	x	DDI, DC	√	√
DSpace	√	QDR, MARC, MODS	√	√
ePrints	x	-	√	√
EUDAT	√	DC, MARC, MARCXML	√	√
Figshare	x	DC	x	√
Zenodo	√	DC, MARC, MARCXML	√	√
Islandora	√	DC, DDI, MODS, METS	√	√

DDI: Data Documentation Initiative; DC: Dublin Core; QDR: Qualification Dataset Register; MARC: Machine-Readable Cataloging; MARCXML: Machine-Readable Cataloging XML; MODS: Metadata Object Description Schema; METS: Metadata Encoding and Transmission Standard

2.5 User interface

Interaction between the user and the software.

Table 5: Comparison for the user interface of component

	User interface	
	Customization of design	Design for mobile
CKAN	√	√
Dataverse	√	√
DSpace	√	√
ePrints	√	-
EUDAT	x	-
Figshare	-	-
Zenodo	√	-
Islandora	√	√

2.6 Articulation with other services

Possibility of platforms embedding in themselves raw data analysis functionalities, through additional plug-ins.

Table 6: Comparison for the articulation with other services component

	Articulation with other services				
	Media viewing and reprodution	Tabular data graph	Georeferenced data analysis	Diverse data types	Data access via API
CKAN	-	√	√	√	-
Dataverse	-	√	√	√	√
DSpace	x	-	√	√	x
ePrints	√	-	-	√	√
EUDAT	√	√	√	√	√
Figshare	√	√	-	√	√
Zenodo	-	√	-	√	√
Islandora	√	√	√	√	√

3 Conclusion and future work

The analysis of the repositories proved that the selection of a platform for data management can be a difficult task, as it is necessary to assess the concrete needs in each situation. An important aspect to focus on is the fact that repositories tend to be increasingly prepared to be in line with data interoperability and accessibility guidelines. Of course, data sharing is one of the goals of data repositories, however, it is necessary to ensure that, when accessed, the data are properly interpreted, as this will guarantee their reproducibility. Data reused by third parties guarantee credit to authors, through citations and references.

It is important to emphasize that repositories are not mere guardians of data. The competent description needs to be promoted through varied metadata models and domain-specific vocabularies. It is necessary to guarantee different conditions of access, such as full access to data or only access to metadata, but with the safeguard of the possibility of contacting the authors. Authors must be safeguarded by associating a DOI (Digital Object Identifier) to the data while facilitating the citation and reuse process. There are several challenges, however, the advantages will outweigh all efforts.

Acknowledgements Joana Rodrigues is supported by research grant from FCT - Fundação para a Ciência e Tecnologia: PD/BD/150288/2019.

References

1. Amorim, R.C., Castro, J.A., Silva, J.R., Ribeiro, C.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society **16**(4), 851–862 (2017). https://doi.org/10.1007/s10209-016-0475-y.
2. Armbruster, C., Romary, L.: Comparing repository types: Challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. SSRN Electronic Journal (2010). https://doi.org/10.2139/ssrn.1506905
3. Guedj, D., Ramjoué, C.: European commission policy on open-access to scientific publications and research data in horizon 2020. Biomedical Data Journal **01**, 11–14 (2015). https://doi.org/10.11610/bmdj.01102
4. Heidorn, P.: Shedding light on the dark data in the long tail of science. Library Trends **57**, 280–299 (2008). https://doi.org/10.1353/lib.0.0036
5. Lagoze, C., Sompel, H., Nelson, M., Warner, S.: The open archives initiative protocol for metadata harvesting (2002)
6. Lynch, C.A.: Institutional repositories: Essential infrastructure for scholarship in the digital age. portal: Libraries and the Academy **3**(2), 327–336 (2003). https://doi.org/10.1353/pla.2003.0039
7. Nelson, M.L., Sompel, H.V.d., Warner, S.: Advanced overview of version 2.0 of the open archives initiative protocol for metadata harvesting. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries (2002). https://doi.org/10.1145/544220.544367