# Comparing Open Data Repositories

**Pedro H. M. Costa**[1], **André F. R. Cordeiro**[1], **Edson OliveiraJr**[1]

[1]State University of Maringá (UEM), Maringá, PR – Brazil

phmc_1995@hotmail.com, cordeiroandrefelipe@gmail.com, edson@din.uem.br

***Abstract.*** *Open Data is one of the main concepts of Open Science, which has the purpose to make scientific research artifacts accessible for everyone. Open data provides recommendations and practices to get access and use data from scientific researches, in a free, permanent, citable, auditable and interchangeable way. To facilitate the data management, it is important to store them in a repository. Considering this context, this paper provides a comparison among five known open data repositories. We performed the comparison taking into account a set of criteria, such as, data format constraints, digital identifier, versioning of published datasets, curators of data collections, metadata schema, versioning and exportation, storage limit, paid services, redundancy and preservation, access controls and APIs. We present results and discussions, in terms of such criteria.*

## 1. Context

Open Science can be explained as a movement towards the sharing of artifacts developed in scientific research. Among such artifacts are processes, codes, experimental packages and articles [Mendez et al. 2020].

Different branches of Open Science can be found in the literature. A taxonomy of Open Science concepts (Figure 1) is presented by the Foster Open Science initiative [Open Science 2020, Pontika et al. 2015]. Each concept focuses on different artifacts or activities. One of its main concepts is Open Data, which aims at managing one or more data sets, in terms of definition, standards, use, reuse, sharing and distribution [Open Science 2020, Mendez et al. 2020]. More examples of concepts can be observed in figure 1.

Figure 1 presents many concepts related do Open Science. For instance, Open Access is associated with the access free of costs of all scientific content. Open Reproducible Research describes the possibility of offer free access to experimental artifacts. Open Science Tools refers to the tools that can help in adoption of Open Science. In figure, it is possible to observe a hierarchical representation of concepts. Each concept can be dismembered in other sub concepts. There are definitions for concepts and sub concepts. More information can be found in [Open Science 2020].

In Open Data, certain characteristics are expected from data produced during scientific research. The data should be accessed, copied, and distributed with no or minimal restrictions. In this context, principles can guide the preparation and availability of any kind of data. An example is the FAIR[1] principle. According to its principle, the data should be Findable, Accessible, Interoperable and Reusable.

[1]https://www.go-fair.org/fair-principles

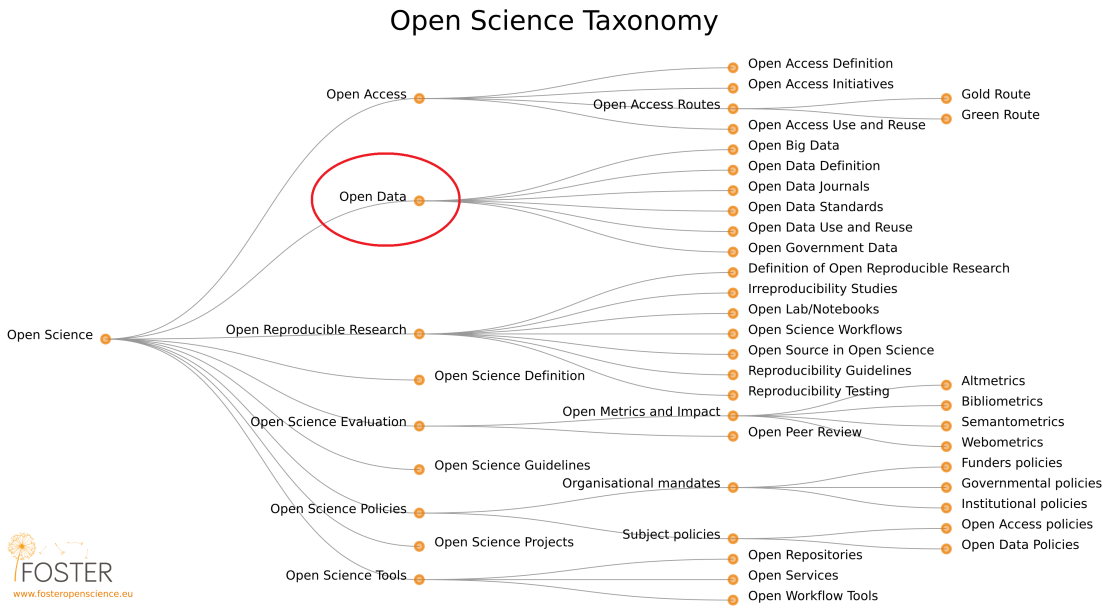


**Figure 1. Foster Open Science taxonomy (adapted from [Open Science 2020])**

The distribution of scientific data favors the achievement of important results, such as, visibility, research partnership, and opportunity to obtain resources [McKiernan et al. 2016]. Even with a few reports on open data, it is possible to observe that scientific data are not commonly shared, at least in specific contexts. [Furtado 2018], for example, presents an analysis of 211 controlled experiments and quasi-experiments carried out in the software product line context. Such analysis was guided by different research questions, such as for documentation, experimental package, and experimental design. It was observed that only 64 (30.3%) experiments explicitly informed about data availability.

The result presented by [Furtado 2018] suggests that the management of scientific data should be approached broadly, considering different stages. One of these stages is data preservation, which involves data storage [Mosconi et al. 2019]. Reliable repositories should be selected for data storage, considering data can be preserved for a long period, or even indefinitely.

Considering the influence of repositories in the context of Open Data, this paper presents a comparison among Zenodo[2], arXiv[3], CiteSeerX[4], UK Data Archive[5] and Figshare[6] repositories, in terms of a set of criteria.

Next sections present more details. Section 2 presents related studies with open data repositories. Section 3 presents the methodology applied in the study. Section 4 presents results and discussion. Section 5 presents final remarks and directions for future works.

[2]https://zenodo.org/
[3]https://arXiv.org/
[4]https://citeseerx.ist.psu.edu/index
[5]https://www.data-archive.ac.uk/
[6]https://figshare.com/

## 2. Open Repositories

Open repositories can be defined as structures that host artifacts and allow free access to them by anyone [Open Science 2020]. Articles and data are examples of artifacts that can be hosted in repositories. In the context of Open Science [Ku and Bao 2017], it is possible to find studies related to open access repositories or open data repositories [Yrivarren 2021, Iddriss and Al Sarraj 2019, Ali et al. 2018, Komiyama and Yamaji 2017, Charalabidis et al. 2016a].

In [Yrivarren 2021], results about the creation of repositories with support for open access in Peruvian Universities and Research Institutes are presented. [Iddriss and Al Sarraj 2019] present results on the availability of open access repositories in West Africa.

In [Ali et al. 2018], the growth in the number of repositories is shown. The repositories considered are registered in the Open Access Repositories Directory (OpenDOAR)[7]. In [Komiyama and Yamaji 2017], the development of a Research Data Management (RDM) is presented. The development project considered the Open Science Framework (OSF)[8]. [Charalabidis et al. 2016a] describe a framework to support the development of open applications. Data and web services are made available.

Open repositories are defined in different environments, whether academic or non-academic [Lima and Peres 2021, Danny et al. 2019, Komiyama and Yamaji 2017, Cheikhi et al. 2012]. In [Danny et al. 2019], the implementation of an open access institutional repository is presented. The repository was developed with the aim of improving the visibility, measurement and impact of scientific publications from the Technological University of Panama. In [Cheikhi et al. 2012], an analysis is presented to identify quality characteristics that can be referenced in the International Software Benchmarking Standards Group data repository (ISBSG).

In [Lima and Peres 2021], a Systematic Mapping protocol for searching and retrieving health applications in non-academic repositories is described. Two guides are presented to assist in using the protocol.

When developing open repositories, different solutions can be used [Medina et al. 2017, Kil et al. 2006]. In [Medina et al. 2017], a strategy for handling open data is presented. Among the main contributions of the work is the analysis of potential benefits of a network of institutional repositories. In [Kil et al. 2006], the OpenArXiv project is described, with a focus on manageability and accessibility. Manageability is related to the application of database techniques. Accessibility is related to the development of an Application Programming Interface (API) to facilitate access to data.

## 3. Comparison of Open Data Repositories

This section presents the methodology adopted for this comparison study as follows.

[7] https://v2.sherpa.ac.uk/opendoar/
[8] https://osf.io/

### 3.1. Goal and Research Questions

This study was carried out to **analyze** open data repositories **with the purpose of** comparing them **with respect to** a set of criteria in **the perspective of** open science researchers **in the context of** known selected repositories.

Considering the characterization of each repository and the established goal, the following research questions were defined:

- **RQ1:** What characteristics are associated with data management?
- **RQ2:** How is metadata managed, in terms of creation, exportation and versioning?
- **RQ3:** How do repositories manage storage and security?
- **RQ4:** What aspects are considered in data access?
- **RQ5:** How to use repository information in software?

The established research questions are important to analyze the obtained results. More details in Section 4.

### 3.2. Planning

We conducted this study taking into account the documentation of each repository, as well as the literature about Open Science and Open Data [Mendez et al. 2020, Mosconi et al. 2019, Medina et al. 2017, Charalabidis et al. 2016b]. After the search for information, we carried out an analysis to characterize each repository. Finally, we answered each defined research questions.

#### 3.2.1. Selection of Repositories

In this study, the selected repositories are Zenodo, arXiv, CiteSeerX, UK Data Archive and Figshare. We chose these repositories based on the Nature's recommended data repositories[9] and on the following papers related to these repositories [Wu et al. 2018, Bodó 2018, Komiyama and Yamaji 2017, Kashireddy et al. 2013, Guild et al. 2010, Kil et al. 2006]. These papers suggest the importance of selected repositories.

#### 3.2.2. Definition of the Comparison Criteria

The selected criteria were cited in one or more repository's documentation. Each criterion is related to the defined research questions presented in subsection 3.1.

**RQ1** has the purpose to analyze characteristics associated with data management. For this question we considered data format constraints, digital identifier, standards in citation and references, versioning of published datasets, and communities and curators of data collections.

**RQ2** is important to understand the metadata context, in terms of schema, versioning and exportation.

[9] `https://www.nature.com/sdata/policies/repositories`

**RQ3** considers characteristics related to data storage and security, storage limit and history, paid services, redundancy and preservation, and access controls.

**RQ4** expresses details on private data or project and licensing.

**RQ5** is focused on software, in terms of the API and existing plugins.

### 3.3. Operation

To conduct this study we:

1. considered different studies to choose the repositories;
2. verified the documentation of each chosen repository;
3. analyzed the structure of the repositories;
4. filled in a comparison table (Table 1) formed with our defined criteria;
5. organized results and discussed them in terms of each defined research question.

Details about the mentioned steps are presented in sequence.

## 4. Results and Discussion

Table 1 summarizes data related to the comparison of repositories. Next subsections present specific analysis.

### 4.1. RQ1 - Data Management

In this paper, five characteristics are associated with data management. About data format constraint, we observed only Zenodo and Figshare do not have restrictions. In terms of digital identifier, we found CiteSeerX does not consider a specific identifier. Zenodo, arXiv and Figshare repositories consider the Digital Object Identifier (DOI). The UK Data Archive considers ORCID.

In the case of standards in citation and references, Zenodo considers the Initiative for Open Citation (I4OC). UK Data Archive uses the APA citation format and Figshare uses DataCite. Regarding versioning of published data sets, little information was found. Only Figshare presents general information.

After analyzing the communities and curators of data collections criterion, we observed the existence of curation processes in some repositories. This is the case of arXiv, UK Data Archive and Figshare repositories. For Zenodo and CiteSeerX, no information was found.

### 4.2. RQ2 - Metadata Management

Metadata management can be analyzed considering creation, versioning and exporting. Creation is related to defining schema. We observed that only Zenodo and UK Data Archive repositories have information about schemas.

With regard to versioning, we found most repositories do not report details about versioning metadata. Only arXiv informs that the metadata is not edited after the data is announced. In the context of metadata export, Zenodo, UK Data Archive and Figshare repositories have different formats. Some of the examples cited are Dublin Core and the Data Documentation Initiative (DDI).

## Table 1. Compared repositories and criteria

| Criteria / Repository | Zenodo | arXiv | CiteSeerX | UK Data Archive | Figshare |
|---|---|---|---|---|---|
| Data Format Constrains | None | LaTeX, AMSLaTex, PDFLaTex, PDF, HTML | PDF, ZIP, GZIP, UNIX, PostScript | TXT, DOC, XTHML, PDF, RTF, SAV, DTA, HTML, others | None |
| Digital Identifier | DOI | DOI | None | ORCID | DOI |
| API services | search and download | upload, access data and metadata | access and extract of metadata, tables and images | access the data | access the data |
| Data Storage and Security | CERN's EOS service [10] | Cornell University | Pennsylvania State University | UK Data Service | Amazon Web Services |
| Storage Limit | 50GB per record | N/A | N/A | N/A | Unlimited public space |
| Paid Services | free subscription | free subscription | free subscription | subscription | subscription |
| Plugins | N/A | N/A | search data by author, by title and general | N/A | Bitbucket, GitLab, Overleaf and others [11] |
| Storage History | Items will be retained during the lifetime of the repository | N/A | N/A | N/A | N/A |
| Metadata Schema | JSON Schema | A schema is described | Some fields are presented | QuDEx Schema | Some fields are described |
| Metadata Versioning | N/A | The metadata is not editable after announcement | N/A | N/A | N/A |
| Metadata Exportation | Dublin Core and DataCite | N/A | N/A | Dublin Core and Data Documentation Initiative (DDI) | Dublin Core |
| Access Controls | closed, open or embargoed access | None | None | Register is necessary | Files may be deposited under restricted open, or embargoed access |
| Licensing | Users must specify a license | CC BY, CC BY-SA, arXiv.org perpetual | Crative Commons Attribution NonCommercial ShareAlike 3.0 Unported | open, safeguard e controlled | CC-BY |
| Private Data or Project | It is possible to store data in a private way | N/A | N/A | Some data are safeguard | It is possible to store data in a restrict way |
| Standards in Citation and References | Initiative for Open Citations (I4OC) | Externals and own format | data sets related to citation are presented | APA citation format | DataCite |
| Redundancy and Preservation | Data are stored in replicas | N/A | N/A | ISO 27001 certification | DuraSpace and Chronopolis |
| Versioning of published datasets | N/A | N/A | N/A | N/A | General information is presented |
| Communities and Curators of Data Collections | N/A | community of volunteer moderators | N/A | curation process | curation service |

*N/A = information not available*

### 4.3. RQ3 - Storage and Security

From the characteristics presented in Table 1 as criteria, five are associated with storage and security. When analyzing the data storage and security criterion, we observed some repositories are allocated in institutions, such as: CERN's EOS Service (Zenodo), Cornell University (arXiv) and Pennsylvania State University (CiteSeerX).

Regarding storage limit, we noticed arXiv, CiteSeerX and UK Data Archive repositories do not present detailed information. Zenodo has a maximum storage limit per record. Figshare features unlimited public storage space. In the case of services offered by repositories, we observed that, in general, free storage is offered. For certain institutions or companies, storage must be contracted. This kind of service is offered by UK Data Archive and Figshare for different institutions.

In the context of redundancy and preservation, arXiv and CiteSeerX do not present detailed information. In the analysis of access control, it was possible to see different options. In Zenodo, there are closed, open and embargoed accesses. In the UK Data Archive, it is necessary to contact the authors of the data in some cases. In arXiv and CiteSeerX no access control is established.

### 4.4. RQ4 - Data Access

Access can be analyzed in terms of repository or data. In the context of data, it is important to assess whether the data can be private and the licenses associated with that data. Regarding private data, we observed that in Zenodo and Figshare, data can be accessed in a mode different of open. For arXiv and CiteSeerX, no information was found. In the UK Data Archive, some data is considered to be backed up.

In terms of licenses, we found repositories inform about the expected licenses. This is the case of arXiv, CiteSeerX and Figshare. In Zenodo, it stands out only that the user must specify a license. In UK Data Archive, the license type is not informed. Just inform that the data must be opened, safeguarded or controlled.

### 4.5. RQ5 - Software

Repositories offer APIs for different activities. Examples of activities include researching, submitting, accessing data, metadata, tables and figures. With regard to plugins, the observed results were different. Only the CiteSeerX and Figshare repositories presented plugins for activities and platforms, respectively.

## 5. Conclusion

This paper presents results about a investigation related to Open Data repositories. The analysis of repositories can aid users on decision making regarding a set of criteria and data for each repository.

In addition, a comparison might be of help to the specification and development of repositories or extensions of them using their APIs, according to open science and open data principles. As future research we are planning to:

- investigate other possible repositories and criteria;
- expand the comparison, with more repositories and criteria;

- elaborate a set of guidelines to use the repositories, according to their characteristics;
- elaborate a set of guidelines to create or extend an open data repository;
- develop an extension of the an open data repository for storage of controlled experiments and quasi-experiments with the following capabilities: data provenance, data curation, metadata support, data management plan models and preservation guidelines.

## Acknowledgments

The authors would like to thank CNPq for a 12-month PIBIC scholarship to Pedro H. M. Costa.

## References

Ali, M., Loan, F. A., and Mushatq, R. (2018). Open access scientific digital repositories : An analytical study of the open doar. In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, pages 213–216.

Bodó, Z. (2018). A citeseerx-based dataset for record linkage and metadata extraction. In *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 230–236.

Charalabidis, Y., Alexopoulos, C., Diamantopoulou, V., and Androutsopoulou, A. (2016a). An open data and open services repository for supporting citizen-driven application development for governance. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 2596–2604.

Charalabidis, Y., Alexopoulos, C., Diamantopoulou, V., and Androutsopoulou, A. (2016b). An open data and open services repository for supporting citizen-driven application development for governance. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 2596–2604.

Cheikhi, L., Abran, A., and Desharnais, J.-M. (2012). Analysis of the isbsg software repository from the iso 9126 view of software product quality. In *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, pages 3086–3094.

Danny, M., Madelaine, F., Huriviades, C., and Dalys, S. (2019). Digital institutional repositories, component of open science to disseminate scientific publications: Case repository utp-ridda2. In *2019 7th International Engineering, Sciences and Technology Conference (IESTEC)*, pages 653–658.

Furtado, V. d. R. (2018). Guidelines for evaluating software product line experiments. Master's thesis, Universidade Estadual de Maringá, Departamento de Informática, Programa de Pós Graduação em Ciência da Computação. (in Portuguese).

Guild, K., Farrera, M. P., Martin, R., Almeida, R., Bontozoglou, A., Patel, M., Yang, K., and Callaghan, V. (2010). Student: Scenarios, technologies and users within the digital essex network testbed. In *2010 Sixth International Conference on Intelligent Environments*, pages 338–343.

Iddriss, Z. and Al Sarraj, A. (2019). Exploring trends in open access repositories: The case of higher education institutions in nigeria, ghana, cabo verde, and senegal. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 366–367.

Kashireddy, S. D., Gauch, S., and Billah, S. M. (2013). Automatic class labeling for citeseerx. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 241–245.

Kil, H., Lee, D., and Fisher, J. (2006). Openarxiv = arxiv + rdbms + web services. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06)*, pages 374–374.

Komiyama, Y. and Yamaji, K. (2017). Nationwide research data management service of japan in the open science era. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 129–133.

Ku, L.-P. and Bao, Q. (2017). The open search.org in open science era: A communication platform for everyone building their repositories and using others. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 124–128.

Lima, L. and Peres, L. (2021). Protocolo de mapeamento sistemático para busca de aplicativos de saúde em repositórios não-acadêmicos. In *Anais do I Workshop de Práticas de Ciência Aberta para Engenharia de Software*, pages 7–12, Porto Alegre, RS, Brasil. SBC.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., et al. (2016). How open science helps researchers succeed. *Elife*, 5:e16800.

Medina, M. A., Sánchez, J. A., Cervantes, O., Benitez, A., and de la Calleja, J. (2017). Lod4air: A strategy to produce and consume linked open data from oai-pmh repositories. In *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 1–8.

Mendez, D., Graziotin, D., Wagner, S., and Seibold, H. (2020). Open science in software engineering. In *Contemporary Empirical Methods in Software Engineering*, pages 477–501. Springer.

Mosconi, G., Li, Q., Randall, D., Karasti, H., Tolmie, P., Barutzky, J., Korn, M., and Pipek, V. (2019). Three gaps in opening science. *Computer Supported Cooperative Work (CSCW)*, 28(3):749–789.

Open Science, F. (2020). Open science taxonomy. `https://www.fosteropenscience.eu/foster-taxonomy/open-science?page=6`.

Pontika, N., Knoth, P., Cancellieri, M., and Pearce, S. (2015). Fostering open science to research using a taxonomy and an elearning portal. In *Proceedings of the 15th international conference on knowledge technologies and data-driven business*, pages 1–8.

Wu, J., Kandimalla, B., Rohatgi, S., Sefid, A., Mao, J., and Giles, C. L. (2018). Citeseerx-2018: A cleansed multidisciplinary scholarly big dataset. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5465–5467.

Yrivarren, J. (2021). Circumstantial reasoning : Creation and management of open access repositories in peru. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7.