Crosswalk among Prominent Open Research Data Repositories

Dr.M. Krishnamurthy
Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, India.
E-mail: mkrishna_murthy@hotmail.com

Dr. Bhalachandra S. Deshpande
Faculty, Bangalore North University, Kolar, India.
E-mail: balusnd@gmail.com

Dr.C. Sajana
Library Assistant, ISRO, Bangalore, India.
E-mail: sajana.drtc@hotmail.com

Received March 06, 2021; Accepted July 02, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18I2/WEB18307

Abstract

Open Access is a synergised global movement using Internet to provide equal access to knowledge that once hid behind the subscription paywalls. Many new models for scholarly communication have emerged in recent past. One among them is institutional or digital repositories which archive the scholarly content of an organization. While the concept of Open Access opened new arena for institutional or digital repositories in the form of Open repositories. Likewise, the Open repositories for Research Data Management (RDM) are initiative to organize, store, cite, preserve, and share the collected data derived from the research. There are many multidisciplinary and subject specific open repositories for RDM offering exquisite features for perpetual management of research data. The objective of the present study is to evaluate features of popular Open Data Repositories-Zenodo, FigShare, Harvard Dataverse and Mendeley Data. The evaluation provided insights about the key features of the selected Open Data Repositories and which enable us to select the best among them. Zenodo provides maximum data upload limit. While the major features required by a researcher like DOI, File Types, citation support, licenses, search (metadata harvesting) are provided by all three repositories.

Keywords

Open Data, Big Data: Zenodo, Figshare, Research Data Management.

Introduction

Modern technology based society provide equal opportunities to all the citizens without any kind of bias. In such potential environment, people actively engage in learning, responsibly make use of resources around them and play a vital role in nation building. The first and foremost need here is to make learning resources like journals easily available. Providing open access to resources is the on-point solution to this scenario. Open Access is a boon to the scholars when subscription prices of the scholarly journals have been rapidly rising (Van Noorden, 2013). Especially, scholarly resources are to be made available freely to the research community. With this objective, a number of open-access journals have increased during the recent years. While, the funding agencies have also strongly voiced to provide free access to research which are based on public funds. Currently, the open repositories for Research Data Management (RDM) are one such initiative to organize, store, cite, preserve, and share the collected data derived from the research. There are many multidisciplinary and subject specific open repositories for RDM offering exquisite features to manage research data for long term. The current article aims to presents a comparative study of Open Data Repositories-Zenodo, FigShare, Harvard Dataverse and Mendeley Data.

Open Access (OA) Conceptualization

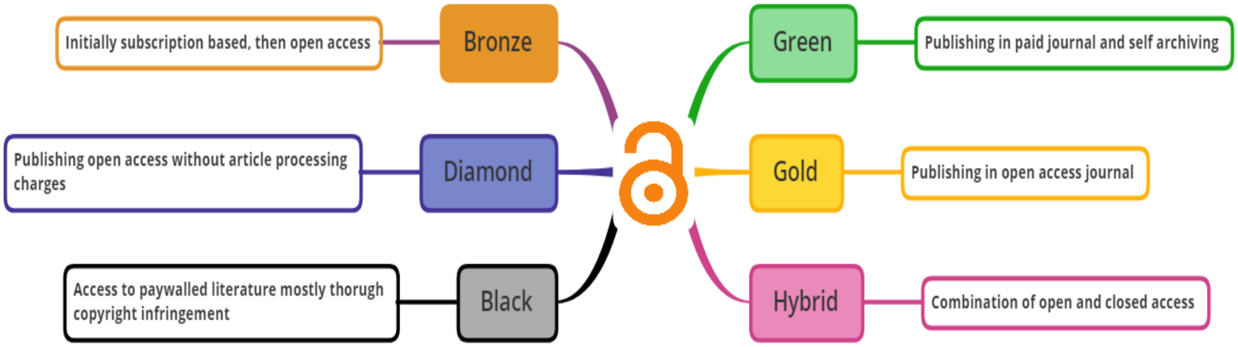
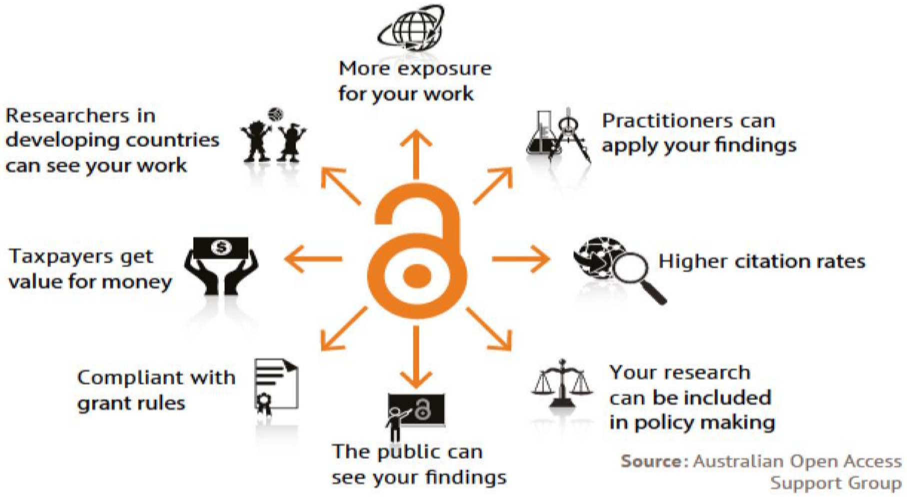

Initially subscription based, then open access
Bronze
Publishing open access without article processing charges
Diamond
Access to paywalled literature mostly thorugh copyright infringement
Black
Green
Publishing in paid journal and self archiving
Gold
Publishing in open access journal
Hybrid
Combination of open and closed access

Figure 1 Types of OA

Budapest Open Access Initiative (BOAI) define “Open Access” as: Open access is making information resources open through Internet for reading, downloading, copying, distributing, printing, searching, or linking to the fulltexts of these e-resources (Van Noorden, 2013).

Another study indicated that freely available articles do have a greater research impact (Antelman, 2004). Metrics indicate that, as of December 2019, Directory of Open Access Journals includes 14,119 open access journals and 4,516,860 Articles from 130 Countries. While in ROAR, currently 4,725 cross-institutional and institutional repositories have been registered. FAIR stands for 'Findable, Accessible, Interoperable and Reuseable', which make the term 'open access' easier to discuss (Wilkinson, 2016).

The interlibrary loan depends on the loaning library in terms of days or weeks to provide the requested article/information. But think of an Open access to the same article available online will be very much fast-paced service for a researcher. The Figure 2 is a direct answer to why we need to support open access.


More exposure for your work
Practitioners can apply your findings
Higher citation rates
Your research can be included in policy making
The public can see your findings
Compliant with grant rules
Taxpayers get value for money
Researchers in developing countries can see your work
Source: Australian Open Access Support Group

Credit: Australian Open Access Support Group
Figure 2 Benefits of OA

Need and Objective of the Study

As aptly said by Sir Isaac Newton that, "If I have seen further it is by standing on the shoulders of Giants", Science is stimulated by exchanging information. When scholars decide to make their data accessible to the public, they are enabling their research to help in ways that go far beyond their own observations. Data exchange has many advantages. It enhances transparency and confidence in their research, make it easier for others to recreate and test their results, and, finally accelerate scientific exploration by encouraging others to re-utilize their data for further research.

The Open repositories for Research Data Management (RDM) helps to systematically share the data with fellow researchers. But, a novice researcher may find it difficult to understand the concept of Open repositories which can be rightly used for managing his/her research data. Next choosing a right Open Repository suitable for researcher’s needs is one another daunting task. Because currently there are many such Open repositories for RDM. Hence, in such contexts the present study provides an overview of Open repositories for RDM and comparison of its major features. This helps in understanding the fundamental concept of Open repositories and selection of an Open repository suiting the need of a researcher. There hasn't been a study that compares prominent Open repositories.

Methodology

The present study uses Descriptive Methodology. It describes the characteristics of selected Open repositories. The researchers first studied various Open repositories and then identified the most widely used Open repositories based on its popularity and mentions on various articles and other sources like websites. The Open repositories for RDM namely Figshare, Zenodo, Harvard Dataverse and Mendeley Data were selected for the study. Then the authors explored the major features of these Open repositories that may be important for its selection by other researchers. These major features were then sorted for any extra texts and only the relevant data is presented in the findings section.

Research Data Management

The publishing of research findings is to reach scientific community. Open access benefits scientists by providing access to articles which they cannot access without subscribing individually or through libraries. Researchers are the direct benefactors from OA, since all journals cannot be subscribed by libraries– this is termed as the "serials crisis" (Laakso & Björk, 2013).

As proposed by its proponents, open access promotes development through research, increase efficiency and also transfer of information (ALA, 2005). Faster discoveries benefit every researcher, not like only for those who are fortunate to access from their library which is subscribing to a specific journal.

Keeping in the view of above scenario, sharing research and its data will promote R&D activities around the world. So, Research Data Management (RDM) aims to organize, store, cite, preserve, and share the collected data derived from the research.

Sharing research data can help to promote your research, increase citation rate and raise researcher. By providing research data, the possibilities that open are:

- Published work can be verified by peers
- Share private links with colleagues and reviewers.
- Control access to data
- Showcase the data with slides and posters
- Backup data and metadata using steadfast cloud based infrastructure
- Securely store research data for long term

RDM platforms help scientists get credit for making their research available. There are many other benefits when research data is made freely available. It also increases transparency and confidence in research findings. Furthermore, data can be reused and speed up the discovery.

Open Access Research Data Repositories

- ### Figshare

Figshare is a web-based interface which is designed for scholarly research data management and dissemination. Figshare supports storing, sharing, discovering research and receive citations for all the research outputs with over 5000 citations of Figshare content till date. Figshare originally was created as a solution to store research outputs in one single place, simultaneously allowing it to be discovered by the academic community. Figshare allows academics upload, cite, share and discover all sorts of research outputs with the secure hosting options and long term preservation of data[8].

- ### Zenodo

The etymology of the word Zenodo is - *Zenodotus* who is father of the earliest recorded usage of metadata and also the first librarian of the Ancient Library of Alexandria. Zenodo supports sharing, cititng, storing and discovering research for all the research outputs. Zenodo hosted by CERN and is funded by EU[7].

- ### Mendeley Data

Mendeley Data enables the researchers to upload and also share their research data. While, the Datasets can be also privately shared among the peers. The Mendeley Data also helps in publishing the data to share with global academic community. Sharing data is significant for science, since Mendeley Data facilitate in reusing the data and play a vital role in encouraging reproducing the research. Also it helps gaining exposure for research outputs, as Mendeley Data provides DOI for every dataset and can be cited[9].

- ### Harvard Dataverse

Harvard Dataverse is open, with a file size limit of 2.5 GB and a dataset size limit of 10 GB. The Harvard Dataverse is an online data archive where researchers can save, upload, cite, and explore their research data. The open-source web application Dataverse, created by Harvard's Institute of Quantitative Social Science, powers the Harvard Dataverse repository. Researchers, papers, and organisations can use Harvard's installation or install the Dataverse web application on their own server. All scientific data from all disciplines is welcome at Harvard Dataverse.

Findings

Features	Zenodo	Figshare	Mendeley Data	Harvard Dataverse
Upload limit	50 GB per dataset	5 GB, 20 GB of free private space	10 GB per dataset.	Up to 2.5GB
Provide DOI	Yes	Yes	Yes	Yes
File Types	Publications, posters, presentations, datasets, images, video/audio files, software and lessons.	Any file format	Images: JPG, tiff, PNG; PDF; Sound; Video; Word; Excel files; Selected programming languages and scripts (e.g.:.json, .java); CSV; Text; and 3D models	R Data, FITS, CSV SPSS, STATA, xlsx
Versions	• Supports versioning on the dataset level. • Once the record has been published, you can no longer change the files in the record, nonetheless, new version can be created (DOI versioning)	• Version control is provided for all publicly accessible data. • Any privately held data may be modified or removed as desired.	Supports versioning of the dataset	Can track version changes.
Software	Zenodo is run with Invenio (an open source software framework), wrapped by a small extra layer of code that is also called *Zenodo*.	No Information	No Information	No Information
Citation Support	Zenodo integrates with GitHub to make software citeable.	No Information	The researchers will be able to cite the study with ease because the data will contain a Force11 compliant citation.	Provides citation for the dataset on website. Allows downloading the citation in several formats.
Servers Support	• OpenStack and the Puppet configuration management system are used to handle Zenodo servers. • The Invenio repository platform programme, which is based on Python and the Flask web development system, are run by Zenodo frontend servers.	To achieve the high security and reliability for the scientific data, Figshare is hosted on Amazon Web Services.	• Data is hosted on Amazon S3 servers for maintaining data confidentiality and security. • Furthermore, released datasets are preserved with Data Archiving and Network Services (DANS) to ensure long-term data protection.	• Glassfish Application Server

Metadata & Search	• For easy and efficient searching, metadata is indexed in an Elasticsearch cluster. • Metadata gets stored in PostgreSQL in JSON format	• Figshare supports OAI-PMH. • Google Dataset Search results provide Figshare objects with the type dataset. These are done by using schema.org JSON-LD markup in the metadata HTML pages of public objects.	The metadata of published dataset is combined to the OpenAIRE portal and to DataCite's metadata index.	Dataverse Network infrastructure provides the flexibility to collaborate, with features like harvesting metadata from one another, creating shared catalogues, and scanning, browsing, and presenting data from multiple archives.
Reporting	After uploading data, Zenodo will take care of the reporting.	Supports reporting	No Information	No Information
Licenses	Allow uploading under a variety of different licenses and access levels	By making data public, it is under the Creative Commons 4.0 licences	Allow to publish it under range of Creative Commons and open hardware and software licences.	The CC0 Public Domain Dedication is given to all datasets attached to the Dataverse registry.
Support for Publishers	No Information	Provide cloud solutions for publishers	Provides a forum for scholars and journals datasets, making them searchable with datasets from over 200 journals.	No Information
API	REST API	open API	open RESTful API	Has many open APIs
Admin Rights	Provide communities to build a hub of curated information with a group of users	• Provide control access to private files and folders with trusted colleagues. • Provide Private link sharing	Supports exchanging unpublished results with colleagues and funding agencies.	Users in a dataverse may be assigned tasks and permissions by the dataverse's administrators.
Certification	No Information	No Information	Has been awarded the industry-recognized CoreTrustSeal seal.	No Information
Institution Support	No Information	Aids academic institutions in storing, exchanging, and handling their entire research output.	Provides organisations with flexible research data management and collaboration solutions, as well as a variety of institutional packages that can be customised to meet particular research data needs.	Any institution may develop a personalised Dataverse set for researchers, departments, and faculty to share their research data using the Harvard Dataverse Repository.
Delete or Edit	No Information	Version control is provided for all publicly accessible data. Any data that is held privately can be modified or removed.	Draft datasets may be deleted via the web interface or API, but released datasets can be removed only after contacting Mendeley Data.	Dataverse can be deleted as long as it is not published. Also, has the option to edit the Dataverse
Peer Review	Promote peer-reviewed openly accessible research, and curate the uploads posted on the front-page	No Information	Datasets that have been uploaded to Mendeley data are currently moderated.	No Information

Discussion and Conclusion

Research data management is a vital scholarly activity required to store and share data for long term. So RDM has to be considered as an important facet in academic publishing. The open research data repositories are supporting researchers in efficient management of research data. Such repositories make data freely available to academic community across the world for reuse during their research project. At this juncture, for an individual researcher it is necessary to understand various open data repositories available.

In this view the current study has compared the features of popular open research data repositories, namely Zenodo, Figshare, Harvard Dataverse and Mendeley Data. The article has elucidated the significance of open access and research data management. Also an attempt has been made to provide general information about the repositories under study. The findings of the study conclude with the understanding that among the three repositories, Zenodo provides maximum data upload limit. The major features required by a researcher like DOI, File Types, citation support, licenses, search (metadata harvesting) are provided by all three repositories. While reporting facility was explicitly mentioned only by Zenodo. Whereas, peer-review is conducted in Zenodo and Mendeley Data.

References

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature News*, *495*(7442), 426. http://doi.org/10.1038/495426a

Antelman, K. (2004). Do open-access articles have a greater research impact?. *College & research libraries*, *65*(5), 372-382.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., & Bourne, P.E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data, 3*. http://doi.org/10.1038/sdata.2016.18

Laakso, M., & Björk, B.C. (2013). Delayed open access: An overlooked high-impact category of openly available scientific literature. *Journal of the American Society for Information Science and Technology*, *64*(7), 1323-1329. http://doi.org/10.1002/asi.22856

ALA Scholarly Communication Toolkit Archived 8 September 2005 at the Wayback Machine. https://blogs.bournemouth.ac.uk/research/researcher-toolbox/research-outputs/bu-open-access-mini-guide/

Zenodo - Research. Shared. (2020). https://zenodo.org/

Figshare - credit for all your research. (2020). https://figshare.com/

Mendeley Data. (2020). https://data.mendeley.com/

Harvard Dataverse. (2020). https://dataverse.harvard.edu/

Piri, Z., Samad-Soltani, T., Elahi, S.M.H., & Khezri, H. (2020). Information Visualization to Support the Decision-Making Process in the Context of Academic Management. *Webology*, *17*(1), 216-226.