Status of Global Research Data Repository: An Exploratory Study

Rashmi Rekha Gohain
G.J. Advani Law College (Affiliated to University of Mumbai), Maharashtra, India, gohain.rashmi@gmail.com

Status of Global Research Data Repository: An Exploratory Study

Rashmi Rekha Gohain
G.J. Advani Law College (Affiliated to University of Mumbai), Maharashtra, India.
Email: gohain.rashmi@gmail.com

Abstract

re3data.org registry is a research data repository and provides information seekers, publishers, libraries and funding organizations an overview of the diverse research data repositories internationally. Under the FAIR Data project and with the 'CoreTrustSeal' certification, re3data is an amiable platform for the researchers to upload and retrieve research data through their appropriate domain repositories. The re3data.org registry of data repository services is explored and relevant data related to general profile, access policies, restriction and licenses, content types, subject coverage and other related services has been collected and analysed in this research study. The study found that, United States has the highest number of data repositories (1102) followed by Germany (433) and United Kingdom (296). India is in 11th position with 51 repositories. Among the repositories, 2059 were disciplinary, 671 were institutional and 291 were of other types. 2574 (42.37%) of the listed institutions were with general responsibility for content development and management of the associated repository followed by 1812 (29.83%) of the institutions as technical host and 1616 (26.60%) as funding institution for the repository. On the other hand, only 1.18% were sponsoring institutions. There was total 135 commercial and 2586 non-profit organisations for the funding of the research data repositories.

Keywords: Data repository, research data, data management, open data, open data repository

Introduction

Research data means representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship (Borgman, 2015). The processing of research data helps in drawing inferences, developing theories or validating original research results. Therefore, it is necessary to ensure long-term preservation and access to valuable research data so that it can be reused by the scientific community when it is required. According to Witt (2012) a number of academic and research libraries are beginning to take a more active role in data management through assisting researchers formulate funder-required data plans, adapting library practice to help organize and describe research datasets, developing data collections and data repositories, digital preservation, and data literacy to find data and integrate it into their learning, teaching, and research.

Research data comprises of both quantitative and qualitative data which are resulted during a research process. The types and format of research data varies among different disciplines including numerals, videos, audios, images, artifacts, etc. Some of the common research data formats are plain text, software application, audio-visual data, structured graphics and text, network-based data, lab notebooks, field notebooks, diaries, questionnaires, transcripts, surveys, codebooks, experimental data, films, photographs, image files, sensor readings, test responses, artifacts, specimens, physical samples, models, algorithms, scripts, content analysis, focus group recordings; interview notes, etc. Considering the heterogeneous nature of research data, it is difficult to store all the data in one data repository. For helping the researchers and research support staffs in selection of data repositories for data sharing and long-term preservation, the "DCC checklist for evaluating data repositories: Version 1.1" was also published by Data Curation Centre in 2015 (Whyte, 2016).

Among the available sources re3data.org is the most comprehensive registry to search and identify data repositories. re3data.org registry was launched in 2012 and it was funded by the German Research Foundation (DFG) between January 2012 to December 2013 and January 2014 to December 2015. The registry was developed under the partnership of Berlin School of Library and Information Science, GFZ German Research Centre for Geosciences, Karlsruhe Institute of Technology (KIT) Library, Purdue University Libraries and German Initiative for Network Information (DINI). To enhance the quality of services through a single, sustainable registry of research data repositories, the re3data.org and Databib.org hosted by Purdue University Libraries were merged in March, 2014. The registry is currently hosted on the web by DataCite and listed about 3595 data repositories across disciplines from all over the world. It enables the researcher to browse data repositories from every domain and in every country by subject, country of origin, or various types of content, and search by any combination of 41 different attributes (Witt, 2018). While selecting a data repository for data submission the researcher/author should try to find out whether the repository provide for free or fees associated with the uploading, maintenance cost like server, cloud storage etc., availability of discovery features using indexing, Search Engine Optimization (SEO) and other discovery tools and access to citation reports, etc.

Review of literature

Antonio et al. (2020) stated that data repositories support qualitative research through secure data management, analysing and sharing among the multi-institutional and geographically dispersed researchers. Akers & Doty (2013) found significant differences related to data management actions, attitudes, needs and interest in support services among the faculty members in different research domain. The two potential services receiving the most interest were faculty workshops on data management practices and assistance with preparing data management plans. Limani et al. (2020) found that there is need for research data curation, preservation, dissemination and access related activities among the researcher in a university system through the institutional repositories which is considered as an important component of a contemporary research infrastructure. Broekstra et al. (2020) stated that the trust of researcher on centralized large-scale data repository depends strongly on whether such data repository benefits the public, the interests of data collectors, the characteristics of the collected data, and application of informed consent for retaining control over personal data.

Kim (2018) based on his study about the contribution of Korea, China, and Japan in data repository stated that the participation of these countries is limited and only 1.8% from China, 3.0% from Japan and 0.3% from Korea are involved in repository building. Hayslett (2015) conducted a study about the different metadata standards used to describe the archived data depending on the discipline of research. The researcher suggested that the researchers can search the metadata standards database of the Digital Curation Centre (DCC) or browse the repository handling datasets of a particular discipline to find the suitable metadata to archive their research data. The author prepared a list of resources associated with data citation, management, finding or acquiring, and archiving/preserving/curating. In another study Kindling et al. (2017) examined the metadata of 1,381 research data repositories listed in the re3data database. It was revealed that the nature of the repositories is heterogeneous depending on the parent institution type, disciplinary background, specialization, access policies and other related requirements.

Rücknagel et al. (2015) outlined the metadata schema of re3data.org which provides the metadata properties about the research data repositories and other optional properties which provides additional information about the data repositories. The author opined that metadata schema helps in “recommending a standard for describing a research data repository; providing the basis for interoperability between research data repositories and re3data.org; and helping data repositories move towards shared standards and practices.” Pampel et al. (2013) outlined the differences between the four repository types i.e., institutional, disciplinary, multidisciplinary and project-specific and tried to describe the features of re3data.org project which helps the researchers to identify the suitable repositories as a producer or user of research data.

Objectives of study

This study has three research objectives and these are as following:

1. To identify and map the research data repositories worldwide;
2. To identify various types of research data repositories on the web; and
3. To find out what licences, software, metadata standards, and various indicators are used by the research data repositories worldwide.

Results and discussion

A case study approach was used, with detailed analysis of the results. The re3data.org registry was explored and data was collected from the registry. The list of data repositories registered on re3data.org was downloaded from the website.

Distribution of data repositories by country

Table 1 list 11 countries with fifty or more number of data repositories registered in re3data.org. Further 250 data repositories are registered as International data repository. Majority of (1102) the repositories registered were from USA followed by repositories from Germany (433), UK (296), European Union (280), and Canada (258) and France (110) with more than hundred repositories. India was in eleventh position with 51 registered data repositories.

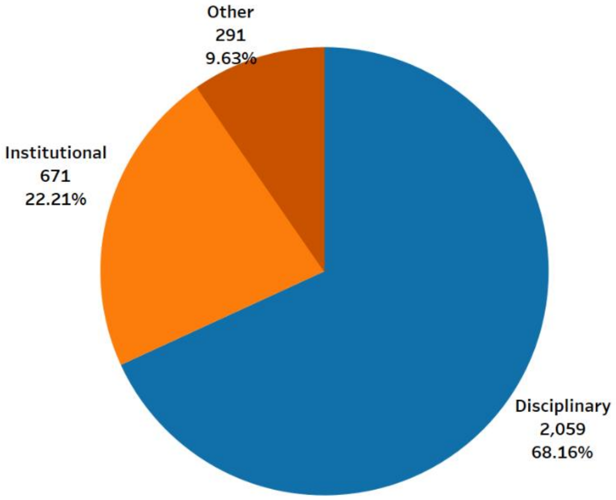
Table 1: Distribution of data repositories by country

Country	No. of data repository
United States	1102
Germany	433
United Kingdom	296
European Union	280
Canada	258
France	110
Australia	92
Switzerland	78
Japan	61
Netherlands	60
India	51
International	250

Data repository types

The research data repositories can be institutional or disciplinary in nature. Institutional data repositories are mostly generic repositories which accepts data from a wide range of disciplines. These institution specific repositories mainly support the universities and research organisations to enhance the visibility of their research output and usually restricted to upload data and documents by their own staff e.g., the Data Repository for the University of Minnesota (DRUM) and Edinburgh DataShare (Banzi, 2019). On the other hand, the disciplinary data repositories include data from specific subject area or discipline. Any research data related to the scope of coverage of the repository can upload in it. The disciplinary data repository could be global, national or institutional in scope. The analysis of re3data.org shows that 2059 (68.15%) repositories are disciplinary/subject specific and 671 (22.21%) are with institutional in coverage.

Figure 1: Types of data repository



Data repositories by responsibility and institution type

While analysing the repositories by institution responsibility types it was found that 2574 (42.37%) of the listed institutions have a general responsibility for content development and management of the associated repository followed by 1812 (29.83%) of the institutions as technical host and 1616 (26.60%) as funding institution for the repositories. On the other hand, only 1.18% were with sponsoring responsibility. Out of total 2721 repositories, 2586 (95.04%) were non-profit organisations while only 135 (4.96%) were commercial organisations.

Distribution of data repositories by AID Systems

Different Author Identification Systems (AID) are used to provide a unique identification number to an author to distinguish and identify the author from other similar or common names (Wagner, 2009). These AIDs have the provisions for individual profiles of the authors, import the list of publications of the author from different citation management tools in the profile, etc.

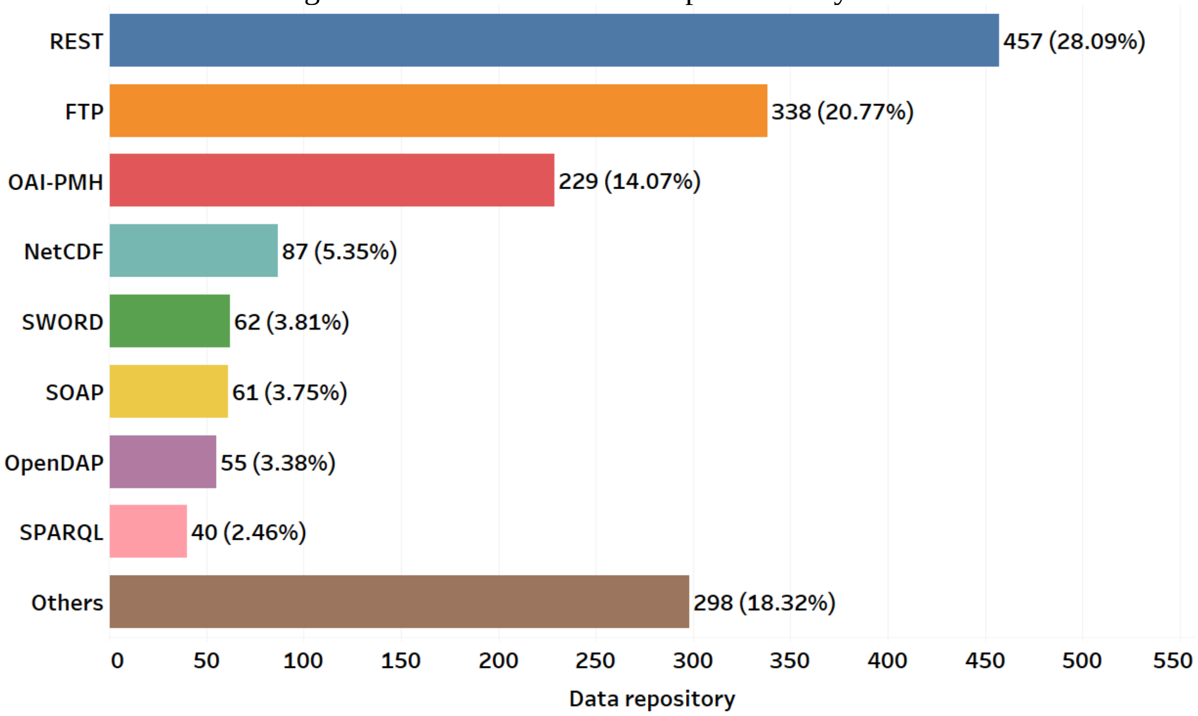
Table 2: Distribution of data repositories by AID

AID Used	No. of repositories (%)
ORCID	193 (33.80)
AuthorClaim	7 (1.23)
ResearcherID	4 (0.70)
ISNI	4 (0.70)
None	352 (61.64)
Other	11 (1.93)
Total	571 (100)

Different AID systems used by the repositories were analysed and it was found that, the most common AID system used is Orchid 193 (33.80%) followed by AuthorClaim with 7 (1.22%) repositories. ResearcherID and International Standard Name Identifier (ISNI) are used by less than 1% data repositories. However, majority 352 repositories were not using any AID system.

An API is an application programming interface and it is a set of rules that allow programs to talk to machines while downloading datasets from the data service provider. Many data service providers have created the API on the data server to help clients to get full datasets or resources from the platform. There are many different ways to get the datasets or resources from the data service provider and the most popular are REST (Representational State Transfer) and FTP (File Transfer Protocol). The analysis of data repositories by type of API (Application Programming Interface) used indicates that REST (Representational State Transfer) is used by 457 (28.07%), followed by FTP 338 (20.76%), OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) 229 (14.06%). On the other hand, APIs such as NetCDF (Network Common Data Form) 87 (5.34%), SOAP (Simple Object Access Protocol) 61 (3.75%), SWORD (Simple Web-service Offering Repository Deposit) 62 (3.80%), OpenDAP (Open-source Project for a Network Data Access Protocol) 55 (3.37%), and SPARQL (SPARQL Protocol and RDF Query Language) 40 (2.45%) were the least used APIs. Further analysis found that 298 repositories used other API.

Figure 2: Distribution of data repositories by API



Distribution of data repositories by certification

Data repositories are certified by the certification organisations considering the wide-ranging characteristics of the repositories developed using internationally recognized standards. CoreTrustSeal is considered as a global framework for repository certification which includes both the extended level (nestor-Seal DIN 31644) and formal level (ISO:16363) of certification of repositories which is valid for three years from the certification date listed within the public application (CoreTrustSeal Standards and Certification Board, 2019). Among the re3data.org listed repositories 104 (38.80%) repositories achieved trustworthy digital repository certification. Further, analysing of the data revealed that other most common certifications were World Data System (WDS) 43 (16.04%) and RatSWD 36 (13.43%), CLARIN certificate B 24 (8.95%), Data Seal of Approval (DSA) 16 (5.97%), DINI Certificate 7 (2.61%), DIN-31644 and Trusted Digital Repository (0.37%). However, 18 (6.71%) of the data repositories were certified with ‘other’ certification.

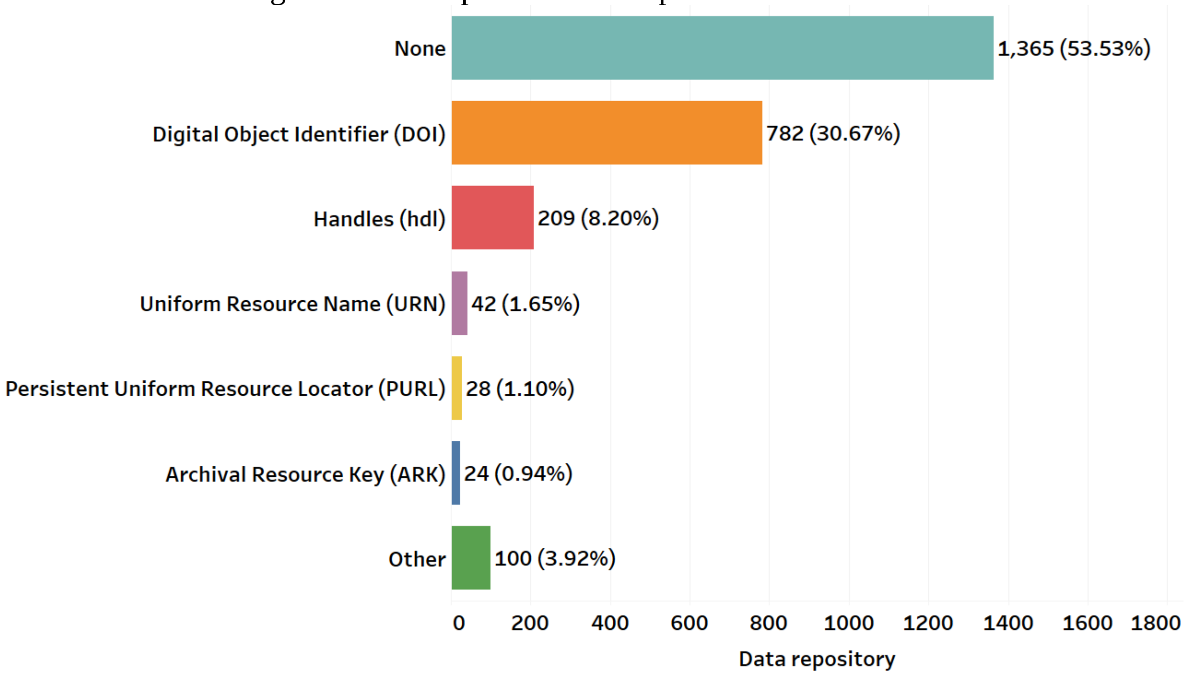
Table 3: Distribution of data repositories by certification

Types of certificate	No. of repositories (%)
CoreTrustSeal	104 (38.80)
WDS	43 (16.04)
DSA	16 (5.97)
RatSWD	36 (13.43)
CLARIN certificate B	24 (8.95)
Other	18 (6.71)
DINI Certificate	7 (2.61)
DIN 31644	1 (0.37)
Trusted Digital Repository	1 (0.37)
Other	18 (6.71)
Total	268 (100)

Data repositories with persistent identifiers

Different repositories assign different Persistent identifiers (PID) for deposited files which are unique by nature. As the URLs are dynamic and changes over time, it is necessary to ensure the retrieve, identity and access to these resources in future. Figure 3 presents the use of different PIDs by the data repositories worldwide. The analysis revealed that the persistent identifiers commonly used were digital object identifiers (DOI) with 782 (30.66%) followed by handles (HDL) 209 (8.19%) and Uniform Resource Names (URN) 42 (1.65%) repositories.

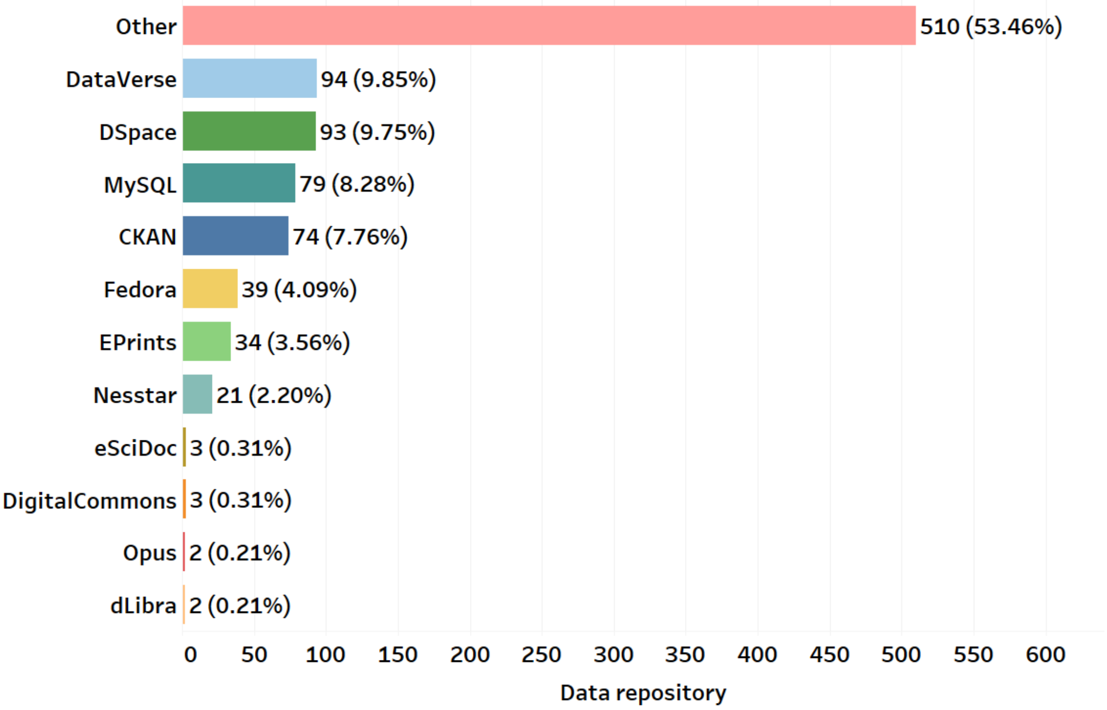
Figure 3: Data repositories with persistent identifiers



Use of repository software

Following Figure 4 shows the data about the use of software in the RDRs. From the analysis it was found that, the software used by majority of the repositories 1231 were unknown and the highest of data repositories 510 are using other type of software which may be developed in-house as per the institution’s requirements. It can also be stated that, DataVerse 94, DSpace 93 and MySQL 79 were the most prevalent repository software followed by CKAN, Fedora, Eprints, and Nesstar.

Figure 4: Data repository software



Metadata standards

It was found that, most used metadata standards were Dublin Core (356) followed by DataCite Metadata Schema (203) and DDI - Data Documentation Initiative (181). The study found that about 28 different metadata standards e.g., ISO 19115, Repository-Developed Metadata Schemas, ISA-Tab, Darwin Care, etc. were used by the data repositories listed in the re3data.org registry (Table 4).

Table 4: Data repositories by metadata standards

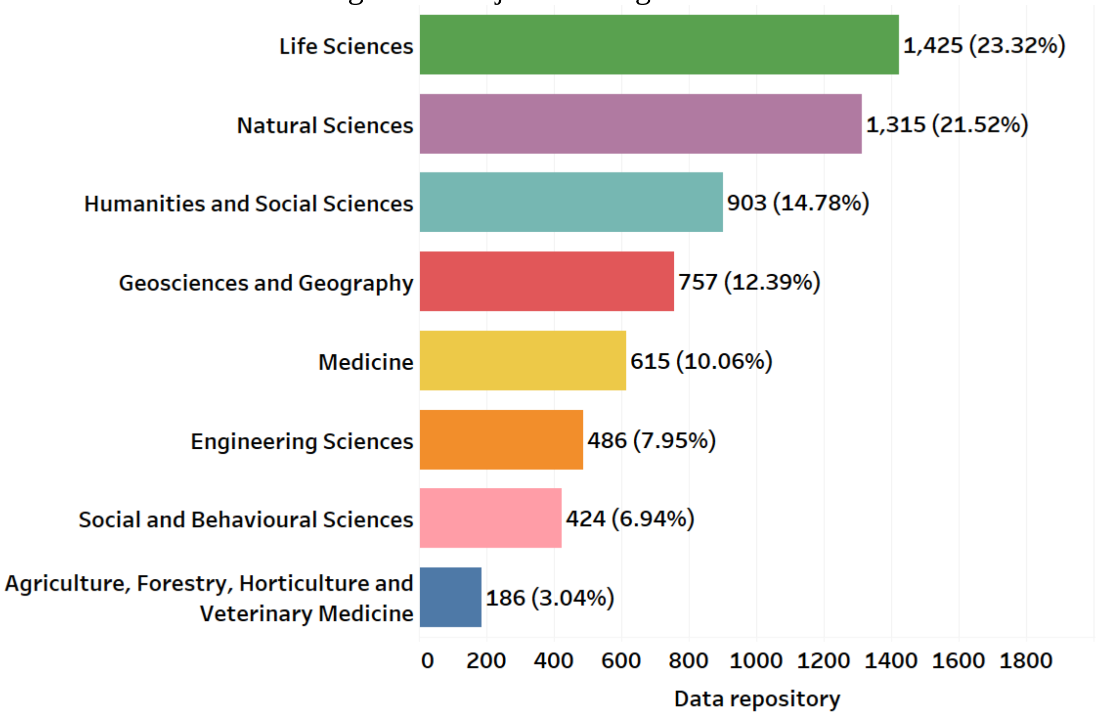
Metadata standards	No. of repositories
Dublin Core	356
DataCite Metadata Schema	203
DDI - Data Documentation Initiative	181
ISO-19115	161
Repository-Developed Metadata Schemas	159
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata	94
DIF - Directory Interchange Format	41
CF (Climate and Forecast) Metadata Conventions	40
EML - Ecological Metadata Language	35
Darwin Core	30
RDF Data Cube Vocabulary	26

OAI-ORE - Open Archives Initiative Object Reuse and Exchange	21
DCAT - Data Catalog Vocabulary	19
ABCD - Access to Biological Collection Data	15
ISA-Tab	13
FITS - Flexible Image Transport System	10
Other	35

Subject coverage and syndications

As far as the coverage of subjects is concerned, majority of the repositories were with collections in the core subjects like Life Sciences 1254 (18.35%), Natural Sciences 1148 (16.80%), Biology 808 (11.82%) and Humanities and Social Sciences 746 (10.91%), as shown in Figure 5. However, some repositories contained datasets on subjects like Medicine 568 (8.31%), Basic Biological and Medical Research 485 (7.09%), Atmospheric Science and Oceanography 382 (5.59%), Social and Behavioural Sciences 378 (5.53%) and Engineering Sciences 369 (5.40%).

Figure 5: Subject coverage

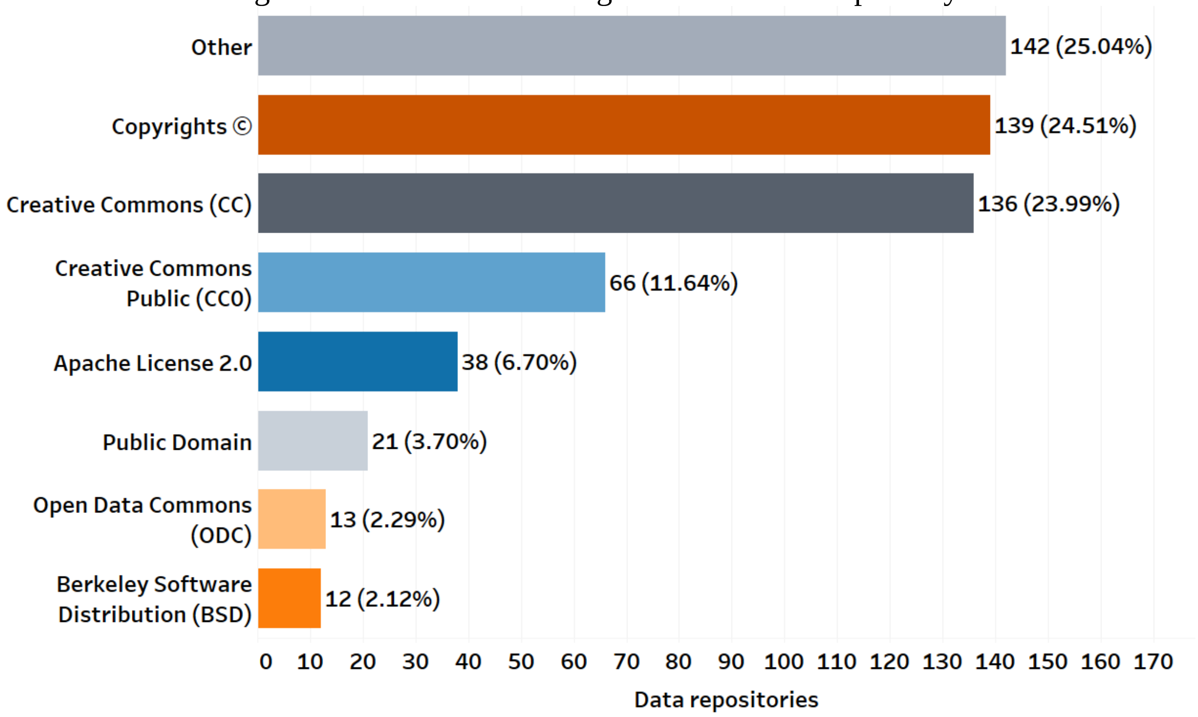


Syndications are used to provide news updates, announcements and other Current Awareness Services (CAS) by the research data repositories. While analysing the data about the syndication used by the data repositories, it was found that, the 528 (81.11%) repositories used Really Simple Syndication or Rich Site Summary (RSS) followed by Atom 121 (18.58%).

Licenses used by data repositories

Figure 6 depicts information about the various types of licenses used by data repositories. Most of the databases 139 (24.51%) are ‘copyright’ compliant and 136 are under Creative Commons licenses which helps to regulate the access and use of the resources by the data repositories. One more variant of Creative Commons license is also used and that is ‘creative commons public (CC0)’ and it is also known as ‘no copyright reserved’ or ‘public domain dedication’. However, about 25.04% repositories used some other database licenses which are not specified.

Figure 6: Licenses used to regulate use of data repository



Data access policy and access restrictions

All the data deposited in repositories were not made open considering the various data protection obligations associated with it. Accessibility of research data is probably the most crucial issue of the data repositories. Open access policy was supported by more than half of the total data repositories, followed by restricted access with 1243 (30.29%) repositories. 367 repositories followed some 'Embargo' period which varied from minimum 6 months to 24 months until which the data remain inaccessible to third party. 'Closed' access policy means external users cannot overcome access barriers and was followed by a smaller number of repositories.

Table 5: Data repositories with access policy and restrictions

Data access policy	Data repositories (%)	Data access restrictions	Data repositories (%)
Restricted	1243 (30.29)	Registration	781 (45.65)
Open	2261 (55.11)	Other	617 (36.06)
Embargoed	367 (8.95)	Institutional membership	125 (7.31)
Closed	232 (5.65)	Fee required	188 (10.98)
Total	4103 (100)	Total	1711 (100)

Different measures adopted for providing access to the data repository. 30% data repositories have restricted users to access the data. It is made mandatory for users to register before accessing the data from repositories and users need to create login username and password with 781 (45.65%) data repositories. 10.98% of the repositories provide access to data on payment of fee. However, institutional membership was required by 7.31% repositories. Although, out of the total 1711 repositories, 617 (36.06%) were using various other mechanisms to restrict the open access to data.

Result reflects the data upload policies used by the repositories. It was found that a significant number of the repositories followed restricted and closed access policy with 1727 (65.85%) and 802 (30.57%) respectively. Further, least number of repositories 94 (3.58%) were with Open data upload policy. Data upload restriction is required to maintain the authenticity and quality of the data. The result shows that Registration 791 (41.25%) was most regularly used mechanism to restrict the upload of research data. It helps to limit the right to upload the data to the registered users only. The analysis further revealed that a significant number of research data repositories require institutional membership 463 (24.14%). However, only 26 (1.35%) repositories charged fees for uploading data.

Conclusion

Research data are of variety of nature and such data can specifically be treated by an information management system like a conventional library or Research Data Repository (RDR). According to Perazzo (2019) Some of major benefits of research data repositories are, researchers can maximize their use of their data; researchers focusing on a particular phenomenon can compare their findings with similar or differing populations and examine changes in a phenomenon over time and data repositories present opportunities to share data and collaborate with other scientists. The RDRs ensures timely access to research data, information exchange and supports decision making, policy formulation, development of products and services. Data repositories helps to find very specific data or data of intrinsic nature (in disciplines like arts and humanities) which are mostly open access. RDR represent an essential stage of summary, abstraction and compression of research data. RDR can be centrally operated i.e., institutional research data repositories and/or locally i.e., disciplinary research data repositories.

The re3data repository is global level registry of research data repositories. Data found in the re3data repository is authorised and validated through producers. It covers research data repositories from all academic disciplines such as Humanities and Social Sciences, Social and Behavioural Sciences, Life Sciences, Medicine, Neurosciences, Agriculture Sciences, Natural Sciences, Engineering Sciences, etc. The re3data repository is comprehensive in coverage and is gaining popularity worldwide due to its bottom-to-up approach for the researchers to store, retrieve and use research datasets, as well as it provides a reliable platform for scientists and information managers. This repository helps researchers to find scholarly institutions, publishers and funding bodies of specific RDR. Its goal is to advocate a culture of increased access, data sharing, and better visibility of research data. The re3data registry follows a unique schema because it is very comprehensive. An editorial team indexes the repositories before it is public. It promotes a culture change of sharing knowledge for better understanding and further services. The re3data repository is somewhat set-up by the researchers and is controlled by the researchers themselves. Thus, the researchers should also archive their research data in open access research data repositories to enable secure access to primary datasets internationally. Researchers can select the repository for deposit by considering the various criteria listed in the re3data.org registry and help significantly in the overall research development of the world.

References

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2), 5-26. http://ijdc.net/index.php/ijdc/article/view/8.2.5

Antonio, M. G., Schick-Makaroff, K., Doiron, J. M., Sheilds, L., White, L., & Molzahn, A. (2020). Qualitative data management and analysis within a data repository. *Western Journal of Nursing Research*, *42*(8), 640-648. https://journals.sagepub.com/doi/abs/10.1177/0193945919881706

Banzi, R., Canham, S., Kuchinke, W., Krleza-Jeric, K., Demotes-Mainard, J., & Ohmann, C. (2019). Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials*, *20*:169, 1-10. https://doi.org/10.1186/s13063-019-3253-3

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world.* Cambridge: MIT Press.

Broekstra, R., Aris-Meijer, J., Maeckelberghe, E., Stolk, R., & Otten, S. (2020). Trust in centralized large-scale data repository: A qualitative analysis. *Journal of Empirical Research on Human Research Ethics*, *15*(4), 365-378. https://doi.org/10.1177/1556264619888365

CoreTrustSeal Standards and Certification Board, CoreTrustSeal (November, 2019). Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022 (version 2.0). Accessed on February 10, 2021 from https://zenodo.org/record/3632533#.X3NDT7DhU5k

Hayslett, M. (2015). Data world does not lack standards. *Journal of Librarianship and Scholarly Communication*, *3*(2), eP1245. http://dx.doi.org/10.7710/2162-3309.1245

Kim, S. (2018). Global data repository status and analysis: based on Korea, China and Japan data in re3data.org. *International Journal of Knowledge Content Development & Technology*, *8*(1), 79-89. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE07409044

Kindling, M., Pampel, H., Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R. & Scholze, F. (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, 23(3/4). http://mirror.dlib.org/dlib/march17/kindling/03kindling.html

Limani, F., Hajra, A., Ferati, M., Radevski, V. (2020) *Requirements and recommendations for university research data repository: A case study.* In: Piet Kommers, Boyan Bontchev and Pedro Isaías (ed.), Proceedings of the 18th International Conference e-Society, April 2-4, 2020 (pp. 51-58). IADIS Press. https://www.diva-portal.org/smash/get/diva2:1449234/FULLTEXT01.pdf

Pampel, H., Paul, V., Frank, S., Roland, B., Maxi, K., Jens, K., Hans-Jürgen, G., Jens, G., Peter, S. & Uwe, D. (2013). Making research data repositories visible: the re3data.org registry. *PloS one*, 8(11), e78080. https://doi.org/10.1371/journal.pone.0078080

Perazzo, J., Rodriguez, M., Currie, J., Salata, R., & Webel, A. R. (2019). Creation of data repositories to advance nursing science. *Western Journal of Nursing Research*, 41(1), 78-95. https://doi.org/10.1177/0193945917749481

Rucknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Reuter, E., Semrau, A., Kindling, M., Pampel, H., Witt, M., Fritze, F., van de Sandt, S., Klump, J., Goebelbecker, H.-J., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Fuchs, C., Spier, S., Kirchhoff, A. (2015). Metadata schema for the description of research data repositories: version 3.0, 29 p. https://doi.org/10.2312/re3.008

Wagner, A. B. (2009). Author identification systems. *Issues in Science and Technology Librarianship*, 59. http://www.istl.org/09-fall/tips.html

Whyte, A. (2016). Where to keep research data: DCC checklist for evaluating data repositories: version 1.1, Digital Curation Centre, (2016). Accessed on January 8, 2021 from http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data/where-keep-research-data

Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration*, 52(2), 172-188. https://doi.org/10.1080/01930826.2012.655607

Witt, M. (2018). 2,000 Data repositories and Science Europe's framework for discipline-specific research data management, (February 13, 2018). Accessed on February 12, 2021 from https://doi.org/10.5438/jeag-2v54