Landscape of Metadata Schemas for Research Data Repositories: FAIRsharing Analysis

Manu T R[1] and Nabi Hasan[2]

[1]Assistant Librarian, Central Library, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India

[2]Head Librarian, Central Library, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India

Abstract

Research data is the recorded information generated while conducting research, writing an article, theses and dissertations, and other research processes. Providing access to research data is a challenge for all stakeholders in the research community. Metadata schemas contain metadata properties describing a research data repository, such as general scope, content, infrastructure, technical, quality, and metadata standards. Several metadata schemas are available for describing the research data repository. However, to facilitate the selection of an appropriate metadata standard for the research repository, the RDA Metadata Standards Directory, re3data.org, and FAIRsharing have compiled a list of metadata schemas in a single platform. FAIRsharing is a curated platform for information and education resources on data and metadata standards, inter-related to databases and data policies. Therefore, researchers preferred FAIRsharing to undertake a study on the assessment of the landscape of metadata schemas indexed by the platform with objectives of identifying the list of metadata schemas, studying their growth and development, distinguishing the disciplinary metadata schemas, identifying the country-wise contributions, organizations, funding and government bodies and institutional contributions in developing and actively maintaining metadata schemas, etc. The result found that FAIRsharing has covered over 1600 metadata schemas covering all the major domains of science and technology, medicine, management, arts & humanities, and social science. It has overwhelming organizations to maintain and fund developing the metadata schemas. The maximum of the metadata schemas is attributed by the Creative Commons attribution, GNU General Public License, Open Data Commons Attributions, etc. Overall, the study found it worthy for data curators, metadata creators, data repository developers, policymakers, research data librarians, etc., to select the appropriate metadata schema for the research data repository.

Keywords: Assessment, Metadata Schemas, Online Administration, Research Data Repositories, Response Rate

---

Corresponding Author: Mr Mant T R, E-mail: manutr91@gmail.com

1. Introduction

Metadata is defined as data about data, but in the context of the research data repositories, metadata is a subset of core standardized and structured data documentation that explains the origin, purpose, time reference, geographic locations, creator, access conditions and terms of use of a data collection. It provides information about data and makes it findable, accessible, interoperable, and reusable (Jeffery & Koskela, 2015). Metadata schemas can be used to generate metadata for research datasets, which differ in types of data, data elements and content. The best practices of metadata for research data are to create a data dictionary, create, manage, and document a data storage system, describe the contents of data files, document taxonomic information, maintain consistent data typing, separate data values from annotations and understand the geospatial parameters of multiple data sources (OpenAIRE, 2023). Therefore, a decision about selecting a metadata schema for a research repository has implications for the quality of analysis. There are several metadata schemas accessible for research data repositories, including general and disciplinary metadata schemas and metadata working groups supporting in development, implementation, and use of the right metadata for the right research data (DataOne, 2023). Directories of metadata schemas such as RDA Metadata Standards Directory, re3data.org, and FAIRsharing directory provide the list of metadata schemas that are available for various subjects, research data types, developed country, parent organization or institutions, license and funding details etc. (Mayernik, 2019). However, research data repository developers need an awareness of available metadata schemas for selecting the suitable schema for creating metadata for their research data.

2. Literature Review

Generally, metadata is defined as data about data (Mayernik, 2019). In the context of data management, metadata is a subset of core standardized and structured data documentation that explains the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of a data collection (Eynden et al., 2009). Metadata provides the information about data that makes it findable, trackable and (re)usable (OpenAIRE, 2020). It is commonly used for the discovery, contextualization, and detailed processing of data (Jeffery & Koskela, 2015) and helps promote researchers' work, better sharing and avoid duplicate research (Wiley, 2014).

It can be characterized according to the attributes of the object, and it leads to different types like descriptive metadata, technical metadata, structural metadata, preservation metadata, provenance metadata, and rights metadata (Treloar & Wilkinson, 2008; Kethers et al., 2010; Kosinov et al., 2019). Structured metadata is the core of data documentation, filling and preservation of research data (Ensom & Corti, 2012). Students must have clear and shared expectations to document and organize the lab notebooks and electronic data files (Carlson & Stowell-Bracke, 2013).

The understanding of data in both immediate and contextual, what data should be collected, what metadata is captured and what discovery services should be established are essential aspects in the metadata

description of research data. The proper metadata must be captured or created to describe the data to support functions such as discovery, use, preservation, and administration (Witt, 2008). The data metadata schemas like:

DataCite, DCAT (Data Catalog Vocabulary), Dublin Core, CERIF (Common European Research Information Format), MARC (Machine-Readable Cataloging), EML (Ecological Metadata Language), FITS (Flexible Image Transport System), MIBBI - Minimum Information for Biological and Biomedical Investigations, Open Archives Initiative Object Reuse and Exchange, Data Package, INSPIRE and DEDI (Data Documentation Initiative)" are being commonly used to describe the metadata of research data. Each metadata schema contains a set of suggested elements or fields (GEO, 2015).

RDF/OWL/SPARQL, W3C's DCAT, schema.org, and OAI-PMH standards are being used (Cudre-Mauroux, 2020; Douglass et al., 2014; Tenopir et al., 2020; Whitmire, Boock & Sutton, 2015) that make data accessible and discoverable so that researchers can find and access, discover, exchange, reuse, combine and share any relevant data (Treloar & Wilkinson, 2008; Eynden et al., 2009).

Every research data element must include the owner of/responsible for each data, name of data, short description, period of moratorium (in months), type of license, and additional notes (Basoni, Menegon, & Sarretta, 2015). The proper use of metadata ensures that data users can access, use, understand, and process data. The researchers need awareness & training on metadata schemas and creating metadata for their research data. The best practices of metadata for research data are to create a data dictionary, create, manage, and document your data storage system, describe the contents of data files, document taxonomic information, maintain consistent data typing, separate data values from annotations and understand the geospatial parameters of multiple data sources (DataOne, 2020). Wolff, Broneske and Koppen (2021) have developed the metadata schema for learning analytics, which is currently lacking for RDM. It increases the findability and extent of it by adding discipline-specific metadata.

The metadata schema is increasing data understandability and reusability in the different phases of the data life cycle (Thanos & Rauber, 2015) and it defined the several activities in creating metadata standards for different research communities. Metadata identifies the data, and unstructured textual description and ensures controlled discovery beyond disciplinary to describe the relationship between an institute or a project (Treloar & Wilkinson, 2008). Every research data element must include the owner of/responsible for each data, name of data, short description, period of moratorium (in months), type of license and additional notes (Basoni, Menegon, & Sarretta, 2015).

The proper use of metadata ensures that data users can access, use, understand, and process data. Also helps data documentation which includes all elements necessary to access, use, understand, and process, preferably via formally structured metadata based on international or community-approved standards (GEO, 2015). The researchers need awareness & training on metadata schemas and creating metadata for their own research data. the best practices of metadata for research data are to create a data dictionary, create, manage,

and document your data storage system, describe the contents of data files, document taxonomic information, maintain consistent data typing, separate data values from annotations and understand the geospatial parameters of multiple data sources (DataOne, 2020).

3. Objectives of the Present Study

The present study is being undertaken to analyse the landscape of metadata schemas available for research data repositories. The focused objectives are as follows:

- To identify the metadata schemas available for research data repositories,
- To study the growth and development of metadata schemas,
- To study and distinguish the disciplinary metadata schemas,
- To identify the country-wise contributions, organizations, government bodies and institutional contributions to developing and actively maintaining metadata schemas, and
- To analyse the metadata schemas by active and readily available for use, supported record types and domains, recommended licenses attributions etc.

4. Methodology

The researchers have used the FAIRsharing directory as a data source for the study, and through REST API required data, including general data, metadata details, grant data and publication data of the metadata schemas, were extracted as on April 25, 2023. FAIRsharing is a community-driven platform to provides information and resources on metadata standards, data policies, databases, and repositories (FAIRsharing, 2023). The community works together to enable FAIR principles by promoting the value and use of the standards, databases, and policies for research data. The extracted data were cleaned and analysed using online visualization tools (RAWGraphs) and advanced Excel features. The result presented in the graphical and tabular formats fulfils the objectives of the study. The researchers did not consider the other two directories as data sources viz. RDA Metadata Standards Directory and re3data.org due to the absence of regular updates and less coverage of metadata schemas respectively.

5. FAIRsharing

FAIRsharing is an informative and educational service that describes and interlinks community-driven standards, databases, repositories, and data policies to increase guidance to consumers of standards, databases, repositories, and data policies, accelerating the discovery, selection, and use of these resources (Ciric, et al., 2022). As of August 2023, FAIRshaing has over 3,868 records which include 1657 standards, 1209 databases and 167 policies (of which 87 are from journals, 33 from funders, 15 from societies, 14 from projects, 13 from journal publishers and 6 from institutions. It covers Natural Science, Humanities and Social Science, Subject Agnostic and Engineering Science (FAIRsharing, 2023). It's been regularly updated by the FAIRsharing open community.

6. Results

The brief findings of the study found that 1657 metadata schemas developed since the 1900s. These schemas are in various stages, 1309 metadata schemas are active and ready for use, 184 metadata schemas have been deprecated, 58 are unsure of their status and 62 are currently in the development stage and not ready for use. Out of 1678 metadata schemas, only 99 (5.89%) have been recommended by a data policy from a journal, publisher, or funder and the remaining 94.10% (1579 schemas) have not been recommended due to various restrictions.

6.1 Growth and development of metadata schemas

As presented in below Figure 1, there has been a constant growth in the development of metadata schemas since 1995 and a maximum of metadata schemas for research data repositories were developed from 2014 onwards. Agencies like the United Nations (665, 39.61%) and countries like the United Kingdom (328, 19.54%) are leading contributors among over 70 countries who have contributed to the development of metadata schemas. India has 12 metadata schemas in their credit. The maximum metadata schemas cover natural science, life science, biology, health science, biomedical science, engineering science and computer science disciplines. And data analysis, data process, data transformation, chemical entity, biological process, and physiological process domains.

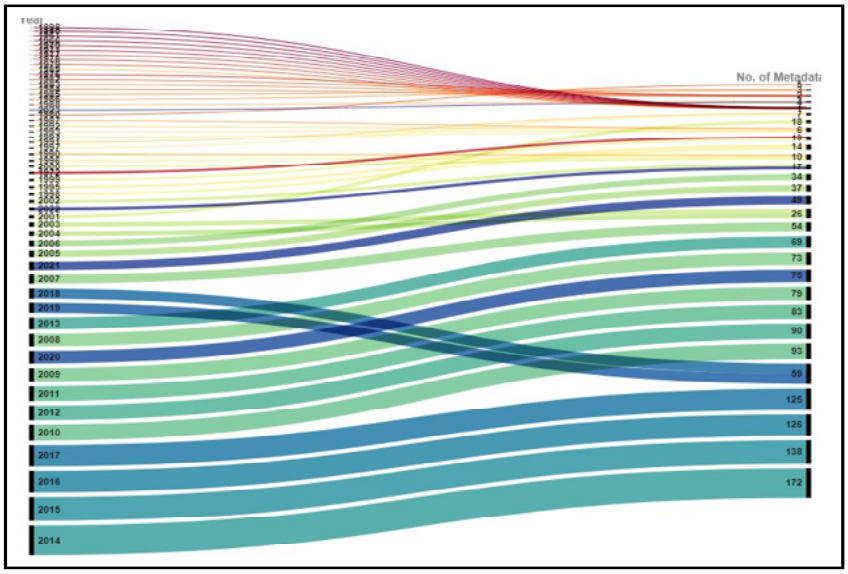

Figure 1: Growth and Development of Metadata Schemas

6.2 Status of metadata schemas

FAIRsharing has covered the metadata schemas available at the various stages, including ready-for-use, deprecated, uncertain and in-development schemas. Over 78.36% of schemas are available for ready-for-

use, 14.48% are deprecated, 3.45% are uncertain and the remaining 3.69% are in the development stage. FAIRsharing is a registry of the metadata schemas on the terminology artefacts, models/formats, reporting guidelines, identifier schemas and metrics. Accordingly, it has 49.76% schema records on terminology artefact, 14.54% are reporting guidelines, 32.95% are model and format, 1.90% are identifier schema and 0.83% are metric relaced metadata schemas. Since FAIRsharing provides the metadata schemas as various stages and record types, it helps the repository developer choose the proper standard.

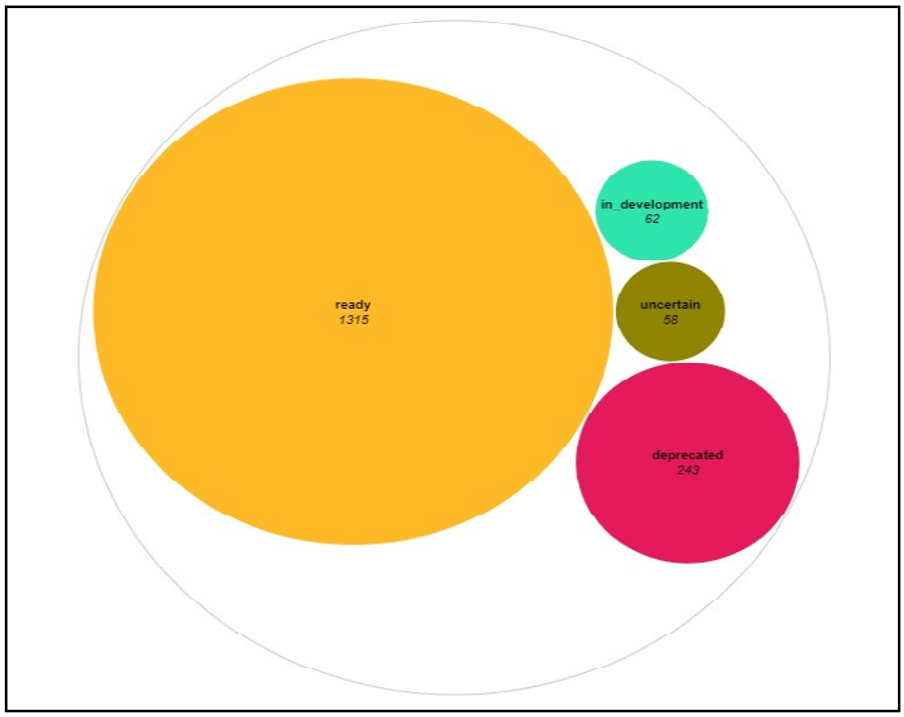

Figure 2: Status of Metadata Schemas

6.3 Subject-wise metadata schemas

The research data is being generated from the various disciplines of the research activities. To create the metadata for multi-disciplined research data, the research data repository developer needs the metadata schemas supporting the multidisciplinary research data. Therefore, assessing the subject area-wise metadata schemas indexed by the FAIRsharing platform is important. Figure 3 graphically presents subject area-wise metadata schemas; as provided there, the maximum of metadata schemas that are available for creating metadata of research datasets are related to Life Science (501), followed by the Biomedical Science (287), Agnostic or unknown (200) and Medicine (92). FAIRsharing platform has classified indexed metadata schemas for over 298 subject areas, including science and technology, Medicine, agriculture, arts & humanities, social science, management, space, etc. Further, it also classified the metadata schemas into domains and taxonomies, which assist in creating the metadata for each of the subjects.

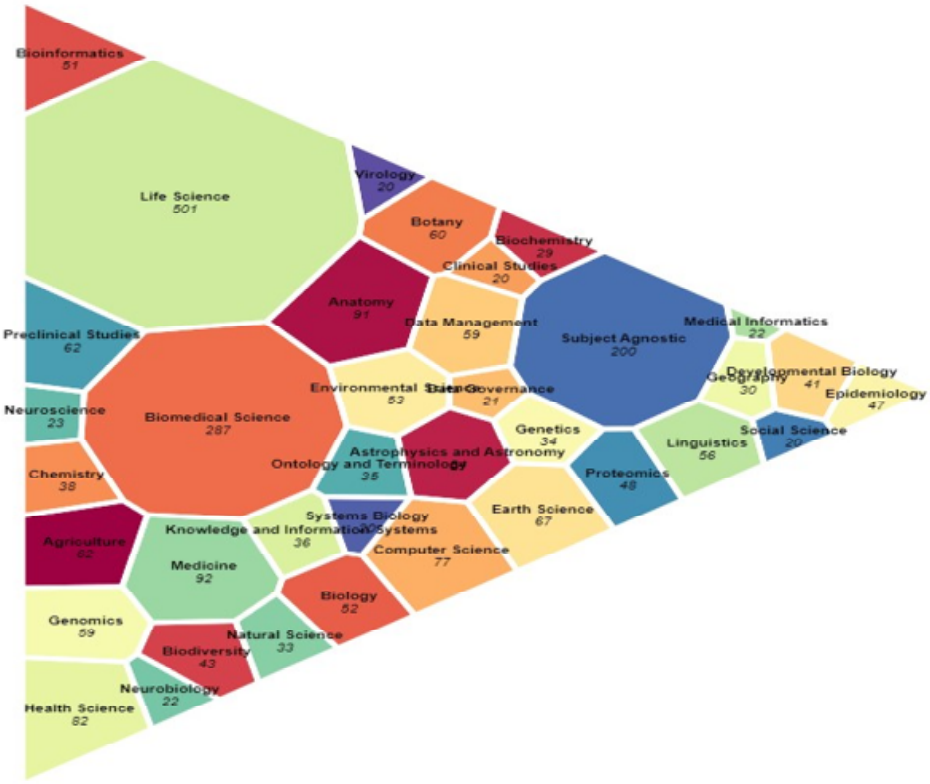

Figure 3: Subject area-wise Metadata Schemas

6.4 Country-wise contribution to metadata schemas development

Several countries have been working on developing metadata schemas for research data repositories according to their requirements and fulfilling their need to create the metadata for research data generated from their organization. FAIRsharing has given the metadata schemas details, including countries that have developed the same. It helps to identify the most contributed countries towards developing metadata schemas. Figure 4 presents the over 96 individual countries that have contributed to the metadata schemas' development. It found the United States has contributed to the development of 658 (39.21%) metadata schemas, 331 (19.72%) metadata schemas as worldwide contributions, United Kingdom (310, 18.47%), France (167, 9.95%), Germany (163, 9.71%) have contributed significantly to the development of metadata schemas for research data repositories. India has collaboratively contributed to developing the 11 metadata schemas, including Synthetic Biology Open Language Visual, Nexus XML, Systematized Nomenclature of Medicine-Clinical Terms, Thesaurus of Plant Characteristics, Breast tissue cell lines, Resource of Asian Primary Immunodeficiency Diseases Phenotype Ontology, Molecular Connections Cell Line Ontology, Common Workflow Language, NeuroML etc. Further, it found that countries have been working collaboratively in developing metadata schemas for research data.

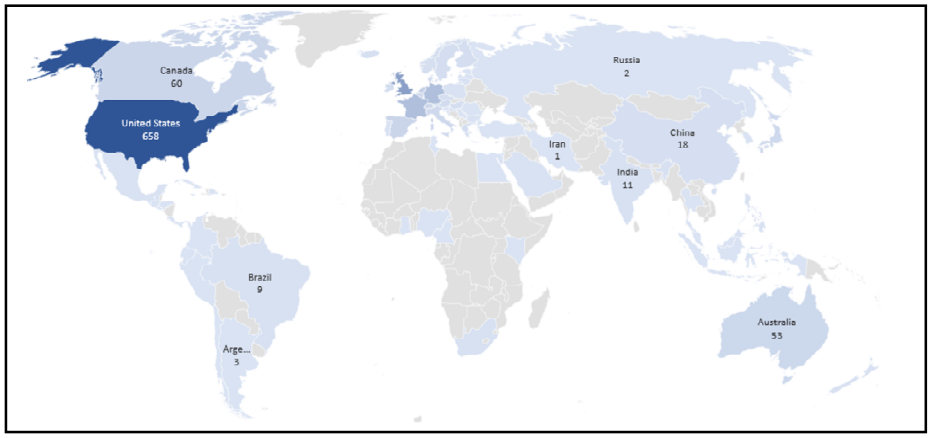

Figure 4. Countries that have developed Metadata Schemas

6.5 Types of Organizations maintaining metadata schemas

There are several organizations and funding bodies that have been involved in the development of these metadata schemas. It found that over 763 organizations have maintained these metadata schemas including prominent organizations such as the National Institutes of Health (NIH), National Library of Medicine (NLM), European Bioinformatics Institute (EMBL-EBI), W3C XSL Working Group etc. These organizations are Consortiums, Government bodies, Research institutes, universities, labs etc. Figure 5 presents the maximum types of organizations contributing to maintaining the metadata schemas. As it showed, over 440 organizations are consortiums, 321 organizations are government bodies, 195 are research institutes, 194 are universities and it includes labs, Charitable foundations, companies, publishers, etc., as part of maintaining organizations.

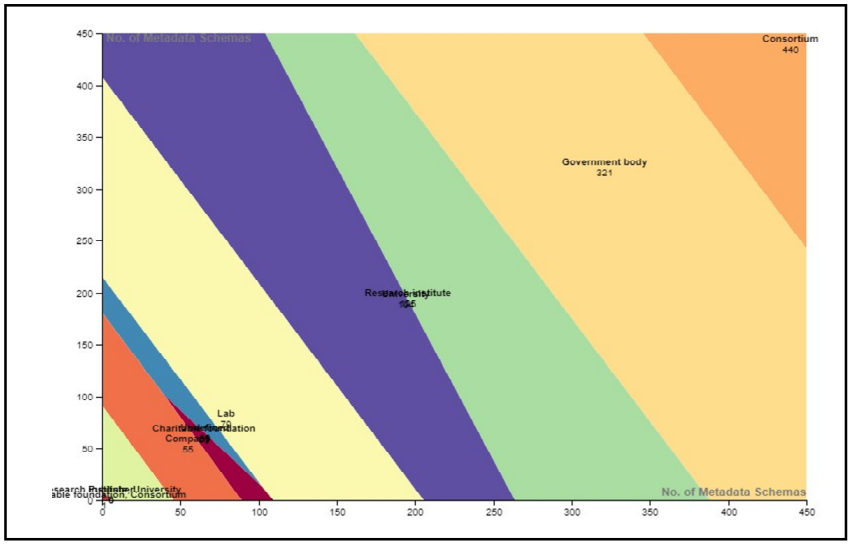

Figure 5: Types of Organizations Maintaining Metadata Schemas

6.6 Organizations funding metadata schemas

Funding is one of the significant aspects of developing any research object. National and international funding bodies have funded research grants to develop metadata schemas for research data repositories. Figure 6 presents such funding bodies that have funded the grant. It found over 82 individual funding bodies have funded research grants to develop the metadata schemas indexed in FAIRsharing.org. The National Institutes of Health (NIH) and National Science Foundation (NSF) are the leading funders who provided maximum grants for developing the 17 and 13 metadata schemas, respectively, followed by the National Institute of General Medical Sciences (NIGMS) and the National Library of Medicine (NLM) who have funded for 8 metadata schemas developments each. The Wellcome Trust, Engineering and Physical Sciences Research Council (EPSRC), European Commission FP7', National Human Genome Research Institute and National Institute of Allergy and Infectious Diseases have given grants for developing the 5 metadata schemas each. It is good to know that international funding bodies have been involved in developing metadata schemas required for research data repositories.

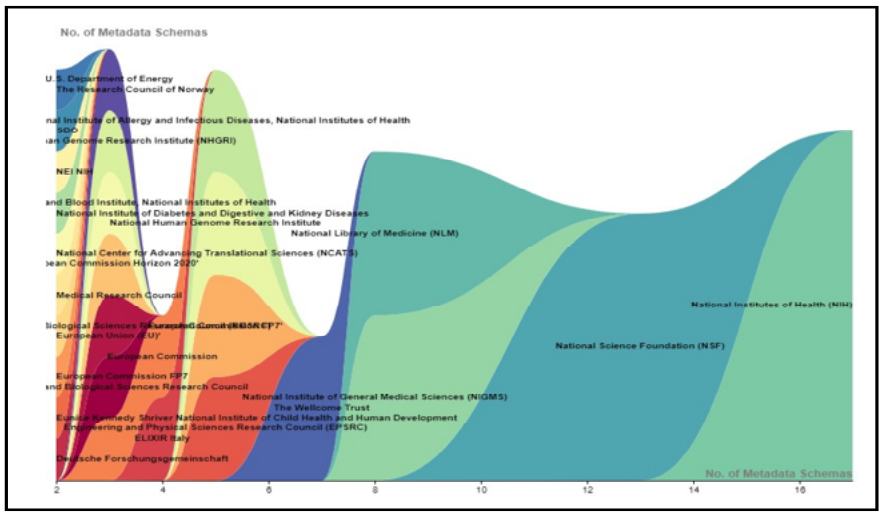

Figure 6: Funding Organizations of Metadata Schemas

6.7 Licenses developing metadata schemas.

Licenses give the freedom to use, reuse, modify, and distribute metadata schemas. So, metadata schemas indexed in the FAIRsharig.org platforms have been attributed with their respective licenses. Therefore, assessing the various licenses attributed in the metadata schemas is found significant. It covers over 139 license agreements, including Creative Commons attributions, MIT license, Open Data license, GNU General Public License, ISO Privacy and Copyright, Library of Congress Legal Information, Open Data Commons, W3C Document License, Apache License 2.0 etc. Figure 7 presents the licenses of metadata schemas; the maximum of metadata schemas supports/recommends the use of licenses attribution of Creative Commons attribution 4.0 international (cc by 4.0) (247, 14.95%) and creative commons cc0 1.0 universal (cc0 1.0) public

domain dedication (74, 4.52%) followed by creative commons attribution 3.0 unported (cc by 3.0) (60,3.81%). Since the maximum of metadata schemas comes under the Creative Commons attributions, therefore, these can be used for reuse and modification on their further term and conditions.

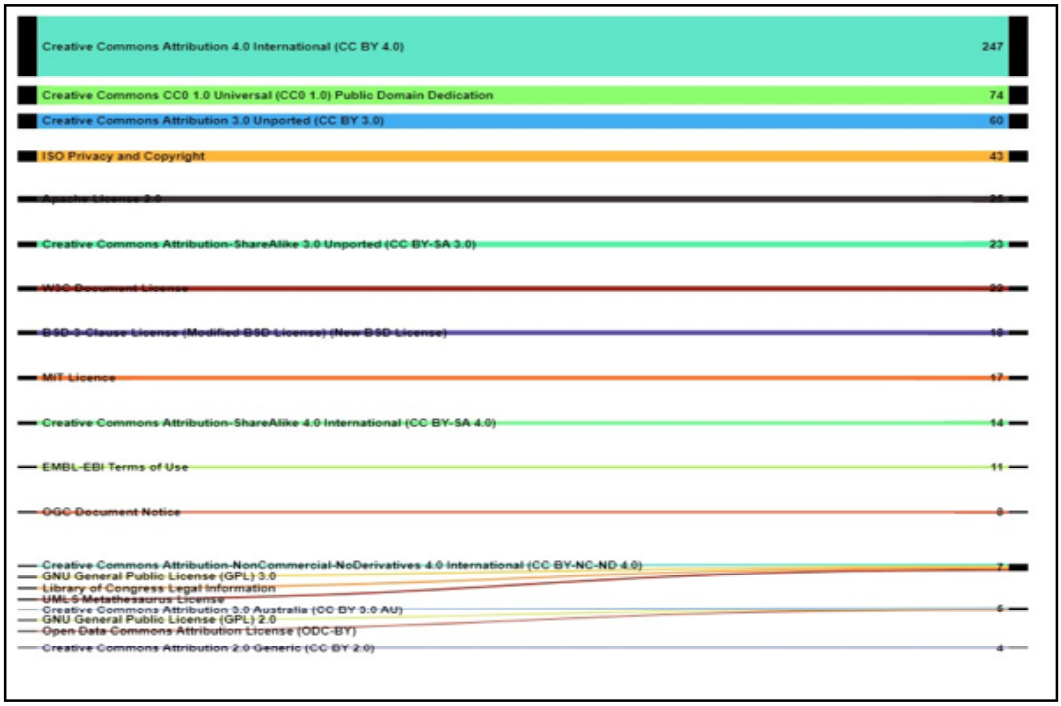

Figure 7: Licenses of Metadata Schemas

7. Conclusion

Overall, the present study gives a clear impression of metadata schemas available for research data repositories. It found over 1500+ metadata schemas with interdisciplinary subjects are available in the FAIRsharing platform and have been used, but 78% were active and ready for use. Therefore, the authors opine that it is essential to ensure that the schema used is stable, applicable for all kinds of data, recommended by the data policies of journals, publishers and, funders etc. The study also helps professionals, decision-makers, and government bodies like India to know which countries are majorly contributing to schemas development, which funding agencies or government bodies have been funding and which organizations or institutions are maintaining the metadata schemas available for research repositories. It also highlights the metadata schemas which recommended using various license attributes. Numerous government bodies, organizations, universities, and institutions are found which have been maintaining, funding, and supporting the development of metadata schemas, including top organizations like the National Institutes of Health (NIH), National Science Foundation (NSF), National Library of Medicine (NLM), European Bioinformatics Institute etc. The study also covered the major top research and academic institutions from India that are developing metadata schemas. Major tags for metadata standardization, institutional repository, research

data, observation, survey etc. are used in the FAIRsharing platform which helps the user retrieve the best available metadata schemas on specific topics.

References

Basoni, A., Menegon, S. & Sarretta, A. (2015). Sailing towards open marine data: the RItMARE Data Policy. ERCIM News (100), pp. 22-23.

Carlson, J. & Stowell-Bracke, M. (2013). Data management and sharing from the perspective of graduate students: an examination of the culture and practice at the water quality field station. portal: Libraries and the Academy, 13(4), 343-361. https://dx.doi.org/10.1353/pla.2013.0034

Cudre-Mauroux, P. (2020). Design considerations on SWITCH's connectome vision. Zorich: A SWITCH Innovation Lab.

DataOne. (2023, 05 18). Metadata. Retrieved from Data Observation Network for Earth: https://www.dataone.org/best-practices/metadata

Douglass, K., Allard, S., Tenopir, C., Wu, L. & Frame, M. (2014). Managing scientific data as public assets: data sharing practices and policies among full-time government employees. Journal of the Association for Information Science and Technology (JASIST), 65(2), 251–262. https://dx.doi.org/10.1002/asi.22988

Ensom, T. & Corti, L. (2012). Research data management at the University of Essex findings from a pilot study. ESSEX: UK Data Archive.

Eynden, V. V., Corti, L., Woollard, M., Bishop, L. & Horton, L. (2009). Managing and sharing data. Colchester: UK Data Archive.

FAIRsharing (2023, 05, 29) FAIRsharing: standards, databases, policies. https://fairsharing.org/

GEO. (2015). Data management principles implementation guidelines. Group of Earth Observations. GEO-XII.

Jeffery, K. G., & Koskela, R. (2015). RDA: the Importance of Metadata. ERCIM NEWS, (100), pp. 23-24.

Kethers, S., Shen, X., Treloar, A.E. & Wilkinson, R.G., (2010). Discovering Australia's research data. Proceedings of the 2010 Joint International Conference on Digital Libraries, JCDL 2010. Queensland. https://dx.doi.org/10.1145/1816123.1816175

Kosinov, A., Erkimbaev, A., Kobzev, G. & Vladimir, Z., (2019). Data curation approach to management of research data. use cases for a upgrade of the thermophysical database thermal. Proceedings of the ÕÕI International Conference DAMDID / RCDL'2019 (pp. 409-419). Kazan: ÀÑÌ Digital Library.

Mayernik, M. S. (2019). Metadata accounts: Achieving data and evidence in scientific research. Social Studies of Science, 732-757.

OpenAIRE. (2023, 04 15). What is metadata for research data? Retrieved from OpenAIRE: https://www.openaire.eu/what-is-metadata

OpenAIRE. (2023, 05 19). What is metadata for research data? Retrieved from OpenAIRE: https://www.openaire.eu/what-is-metadata

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLOS ONE, 15(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Thanos, C. & Rauber, A. (2015). Scientific data sharing and re-use. ERCIM News (100).

Treloar, A. & Wilkinson, R. (2008). Rethinking metadata creation and management in a data-driven research world. 2008 IEEE Fourth International Conference on eScience (pp. 782–789). IEEE. https://dx.doi.org/10.1109/eScience.2008.41

Whitmire, A. L., Boock, M. & Sutton, S. C. (2015). Variability in academic research data management practices: implications for data services development from a faculty survey. Program, 49(4), 382–407. https://dx.doi.org/10.1108/PROG-02-2015-0017

Wiley, C. (2014). Metadata use in research data management. Bulletin of the Association for Information Science and Technology, 40(6). http://dx.doi.org/10.1002/bult.2014.1720400612

Witt, M. (2008). Institutional repositories and research data curation in a distributed environment. Library Trends, 57(2), 191-201. http://dx.doi.org/10.1353/lib.0.0029

Wolff, L., Broneske, D. & Köppen, V. (2021). A first metadata schema for learning analytics research data management. o-bib. Das offene Bibliotheksjournal. 8(4). https://doi.org/10.5282/o-bib/5735