


Spring 2017

# The Current State of Meta-Repositories for Data

Cynthia Hudson-Vitale

*Washington University in St Louis*, [chudson@wustl.edu](mailto:chudson@wustl.edu)

Follow this and additional works at: [https://openscholarship.wustl.edu/lib\\_papers](https://openscholarship.wustl.edu/lib_papers)

 Part of the [Archival Science Commons](#), [Scholarly Communication Commons](#), and the [Scholarly Publishing Commons](#)

---

## Recommended Citation

Hudson-Vitale, Cynthia, "The Current State of Meta-Repositories for Data" (2017). *University Libraries Publications*. 19.  
[https://openscholarship.wustl.edu/lib\\_papers/19](https://openscholarship.wustl.edu/lib_papers/19)

This Book Chapter is brought to you for free and open access by the University Libraries at Washington University Open Scholarship. It has been accepted for inclusion in University Libraries Publications by an authorized administrator of Washington University Open Scholarship. For more information, please contact [digital@wumail.wustl.edu](mailto:digital@wumail.wustl.edu).



## CHAPTER 11

# The Current State of Meta-Repositories for Data

*Cynthia R. Hudson Vitale*

## Introduction

Researchers have many options available to them in order to fulfill individual, funder, and publisher requirements to deposit and share research data. Thus many of their research outputs, including data, may be scattered across various institutional, domain, funder, publisher-supported, and general repositories and websites. Given this, a faculty member searching for data sets similar to his or her own research may find it difficult, if not impossible, to discover relevant sources. Without direct connections among the various research outputs, there are few mechanisms for anyone to understand what data, article, and code are related to the same research. This is a significant scholarly communications issue. Recently, much work has developed around online solutions to federate and link the records across these dispersed repositories, creating large meta-repositories of data.

Traditionally, in the scholarly literature meta-repositories of data have been categorized as digital libraries. What constitutes a digital library is complex, often defined ambiguously by the research community describing it.<sup>1</sup> When the World Wide Web was in its nascent stages, it too was considered a digital library. A more library-centric definition developed in the late 1990s, during which digital libraries were more closely tied to traditional libraries that had collection development plans, ensured the persistence of materials, preserved documents, and distributed the resources.<sup>2</sup> While meta-repositories of data fit this definition, they also have

a number of distinct qualities that set them apart, including a close focus on research materials and the aggregation of metadata or data from dispersed sources.

A previous study of digital libraries that are more similar to these meta-repositories of data, compiled by the Digital Repository of Ireland (DRI), focused on how digital objects were being cared for internationally.<sup>3</sup> The authors found three different models: the metadata aggregator, the single-site repository, and the multi-site repository. DRI also indicated that funding agencies place a greater emphasis on access rather than preservation of the digital content, which may ultimately put the ongoing availability of content at risk.

Extending the work completed by DRI, this chapter comparatively analyzes the major international meta-repositories of data to better understand their goals and missions, overlaps in services and content, and any common challenges.

## Community Initiatives and Solutions to Support Meta-Repositories of Data

Though the scholarly literature around meta-repositories of data is not extensive, a number of international organizations have become more inclusive of data repository agendas by establishing working groups to address repository technical issues, metadata challenges, and interoperability.

Founded in 2009, the Confederation of Open Access Repositories (COAR) seeks to create community and support for repositories worldwide. Current members include the Vienna University Library and Archive Services, the University of Antwerp, McMaster University Library, bepress, and the World Bank, to name a few.<sup>4</sup> The COAR organization and community builds capacity, aligns policies and practices, and acts as a global voice for the repository community. COAR's approximately 100 members represent libraries, universities, research institutions, government funders, and others. According to COAR's 2016–2018 strategic plan, one of its primary objectives is to work towards interoperability with research data management repositories and systems.<sup>5</sup> Interoperability work such as this might allow federated data repositories to more easily aggregate metadata records and exchange information.

The Research Data Alliance (RDA) was established in 2013 as a grass-roots organization that builds the technical and socio-technical infrastructure for data sharing.<sup>6</sup> It is organizationally comprised of approximately sixty-two interest and working groups that include focus on everything from wheat interoperability to sharing sensitive data and developing a data type registry, to name a few. One community group that includes repository interoperability among its goals is the

Repository Platforms for Research Data Interest Group.<sup>7</sup> A deliverable of this group is to create a matrix of functional requirements related to repository platforms, which may also relate to specifications for a generic application programming interface. A newly proposed group, titled Research Data Repository Interoperability, is looking specifically at research data repository interoperability as a working group. The main objectives of this group are to identify, evaluate, and establish standards for interoperability between different research data platforms. Already, repository developers representing DSpace, Hydra, Fedora, DataOne's Metacat, and others have agreed to implement these recommendations upon the close of the working group.<sup>8</sup> These types of community-developed and -initiated projects ensure wide adoption and solutions that fit the needs.

Organizations that support the quality and accessibility of data are not new. The International Council for Science: Committee on Data for Science and Technology (CODATA) is an organization established over forty years ago. One of CODATA's main objectives is to facilitate international cooperation among those institutions collecting, organizing, and using data.<sup>9</sup> This work is primarily facilitated through the committees and working groups focused on projects of specific scope, such as legal interoperability.

Finally, the International Council for Science: World Data System (ICUS/WDS) is a unique organization that promotes universal access and long-term stewardship of quality-assured scientific data and data services products.<sup>10</sup> This organization, comprised of working groups, is unique because it also provides services and aggregates data from member organizations, thereby acting as a "meta-repository."

Meta-repositories of data participate in, support, and are putting into practice many of the recommendations and outputs developed or in development by these community initiatives. Yet understanding how these meta-repositories of data work together, overlap, or complement each other has not been examined. Thus, the goal of this study is to comparatively analyze these systems in order to better understand the current state of meta-repositories for data.

## Methods

A unified term to describe meta-repositories of data currently does not exist, which makes conducting Web searches to identify these systems impossible. Conducting Web searches using the terms *federated repositories* and *repository aggregator* resulted in zero relevant systems. Thus, the meta-repositories of data described here were primarily identified through the author's knowledge of such systems and suggestions from colleagues.

Thirteen meta-repositories were chosen for analysis based upon the following criteria:

1. Content: The meta-repositories of data were receiving or harvesting data (either metadata or digital data objects) from individual repository platforms.
2. Language: The meta-repositories of data websites were written in English.
3. Spatial: International repository aggregators were within scope of this analysis.

The thirteen repositories are listed in table 11.1.

**TABLE 11.1**  
The Meta-Repositories Chosen for Analysis in this Study

Meta-Repository	Mission	URL
1. Australian Research Data Commons (ANDS)	ANDS is a system built and maintained in Australia to <ul style="list-style-type: none"> <li>• “make Australian research data collections more valuable by managing, connecting, enabling discovery and supporting the reuse of this data”</li> <li>• “enable richer research, more accountable research; more efficient use of research data; and improved provision of data to support policy development.”<sup>a</sup></li> </ul>	<a href="http://ands.org.au">http://ands.org.au</a>
2. Beilefeld Academic Search Engine (BASE)	BASE is a portal established by Bielefeld University Library, United Kingdom that integrates Open Archives Initiative (OAI) resources as one information type among others into the local digital library environment, together with catalogs, article databases, and digitized collections.	<a href="https://www.base-search.net/">https://www.base-search.net/</a>
3. Connecting REpositories (CORE)	CORE is a UK-based meta-repository that seeks “to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public.” <sup>b</sup>	<a href="https://core.ac.uk/">https://core.ac.uk/</a>
4. Data.gov	Data.gov is the home of US government metadata. Non-federal data sources can also be added to the data set voluntarily.	<a href="http://www.data.gov/">http://www.data.gov/</a>

**TABLE 11.1** (continued)

Meta-Repository	Mission	URL
5. Data Archiving and Networked Services (DANS)	Developed in the Netherlands, DANS is a service institute that promotes sustained access to digital research data.	<a href="http://www.dans.knaw.nl/en">http://www.dans.knaw.nl/en</a>
6. DataBridge	DataBridge is a cross-institutional collaboration that aims to make the “long tail” of data more discoverable.	<a href="http://databridge.web.unc.edu/">http://databridge.web.unc.edu/</a>
7. DataCite	DataCite is an organization that works with data centers to assign digital object identifiers to research assets.	<a href="https://www.datacite.org">https://www.datacite.org</a>
8. EUDAT	EUDAT is a system that includes data access, deposit, sharing, archiving, identification, and discovery of research data produced across the European Union.	<a href="https://eudat.eu">https://eudat.eu</a>
9. ICSU/World Data System (WDS)	Launched in Japan, ICSU/WDS research data system seeks to enable universal and equitable access to scientific data.	<a href="https://www.icsu-wds.org">https://www.icsu-wds.org</a>
10. OpenAIRE	Initiated in the European Union, OpenAIRE brings together scholarly metadata to support open scholarship and improve the reuse of publications and data.	<a href="https://www.openaire.eu/">https://www.openaire.eu/</a>
11. OpenDOAR	OpenDOAR is a directory of open-access academic repositories.	<a href="http://opendoar.org/">http://opendoar.org/</a>
12. OneRepo	OneRepo is a system that seeks to bring together all open-access scholarly articles.	<a href="http://onerepo.net">http://onerepo.net</a>
13. SHARE	SHARE is a metadata data set about research and scholarly activities through the research life cycle (such as data management plans, funder information, articles, data sets, etc.)	<a href="http://share-research.org">http://share-research.org</a>

a. “About Us,” Australian National Data Service, accessed May 26, 2016, <http://www.ands.org.au/about-us>.

b. “About CORE,” CORE homepage, accessed May 26, 2016, <https://core.ac.uk/>.

It should be noted, that while LaReferencia is a known meta-repository for South America, the website is entirely in Spanish. Although OpenDOAR is a directory of open-access repositories, it also includes a Google search widget that allows a user to search across the content of the repositories it indexes; thus, it was included in this study.

The comparative analysis was conducted by evaluating the websites of each meta-repository of data across fifteen variables in four distinct areas (see table 11.2) with the goal of better understanding the repository aggregator's background, content coverage, metadata employed, and functionality of the search interface. All data for the analysis was manually collected during the period October 10, 2015–April 7, 2016. The author searched primarily through each website's About pages and search interfaces and used white papers and other website documents to collect the comparative data. The raw data, along with hyperlinks to the document where the information was collected from, is available in the data set that accompanies this chapter.

**TABLE 11.2**  
**The Meta-Repository Website Analysis Used Variables**  
**Categorized into Four Areas**

Area	Variables Collected
Background	Date founded, goals/vision, mission, funding model
Content Coverage	Time span, spatial/geographic parameters, domain specificity, data types, providers, number of records, update frequency
Metadata	Standards, elements
Functionality	Faceted searching, feeds/alerts

## Results

The results of the website analysis show various points of similarities among the thirteen meta-repositories of data. Six of the meta-repositories were created to support national missions to ensure quality data and accessibility (meta-repositories 1, 4, 5, 8, 9, 10), while the remaining were created as responses to growing scholarly communication needs, to maximize research impact, and to otherwise promote science. For example, the mission of the ANDS is to make Australian research data collections more available “by managing, connecting, enabling discovery and supporting the reuse of this data.” In contrast, SHARE’s mission is “to maximize research impact by making a comprehensive inventory research widely accessible, discoverable, and reusable.”

All of the repository aggregators analyzed, except for BASE (established in 2004), were established within the last ten years, with the majority ( $n=8$ ) established or founded within the last six years (meta-repositories 3, 4, 6, 7, 8, 10, 12, 13). The repository aggregators fell into four distinct funding categories: those that are federally or nationally funded ( $n=6$ ), commercially and organizationally

funded ( $n=4$ ), grant funded ( $n=2$ ), and one that is currently seeking funding or whose funding is unsecured ( $n=1$ ).

## Content

From a content perspective, the majority of the meta-repositories were harvesting content from repositories worldwide ( $n=9$ ), while two were limited to nations and one was unknown in spatial coverage. None of the repository aggregators were limited to a specific domain (i.e., gathering source information only from a specific scientific discipline). While all of the repository aggregators had metadata about data sets in their systems, many also had articles, theses and dissertations, and conference papers and presentations. One repository aggregator, OpenDOAR, also included content such as audiovisual material and learning objects. Most systems were simply aggregating the metadata, but a handful of the meta-repositories had the actual digital asset stored, including CORE and Data.gov.

The number of providers varied significantly among the meta-repositories, ranging from 20 (OneRepo) on the low end to over 6,000 on the upper end (CORE). There was a low, but surprising, amount of overlap found among the institutional repositories covered within these systems. Of the thirteen repository aggregators, only five made their provider list available. Of these, over 1,400 repositories were aggregated overall. While no deduplication was completed as part of this analysis, these aggregators have collectively brought together millions of records. Individually, some of the aggregators did not release how many records they had (OpenDOAR and OneRepo). Time spans of the content found in the aggregated repositories were also difficult to determine. BASE had the longest known temporal span, with records available for materials created in the 1000s.

## Functionality

In regard to search features and faceting found in the meta-repositories for data, all working systems had some type of advanced search limiters. The most common types of features were facets that allowed the user to limit the results by a subject area, institution, or publication year. The Australian National Data Service had a unique function that allowed the user to find related people and related organizations from a search query.

Conducting a search, having appropriate results, and accessing the data set are a primary goal of these systems, but being able to download the metadata of the search results or export metadata in some manner was investigated as well. Seven of the meta-repositories for data had a tool to allow the user to export search results or access the underlying metadata for records in the repository. These tools ranged in implementation from e-mail alerts and SPARQL endpoints



to more robust APIs. Two of the systems, SHARE and OpenAIRE, were developing alerting tools for searches. These tools, such as, SHARE Notify,<sup>11</sup> allowed the user to conduct a search across the SHARE data set and set up an atom feed to receive real-time updates. Use cases for this tool are many, but include the ability to stream this data to a web interface that would keep researchers up-to-date on relevant scholarship or alert local institutional repositories about new faculty-created materials available for harvesting. The Literature Broker Service, in development at OpenAIRE, is similar to the latter Notify use case. It is a subscription-based system that aims to support institutional repository managers by altering them to new publication objects not currently in their collections. This system has the added benefit of disseminating additional or updated metadata related to records already in the repository.<sup>12</sup>

## Metadata

One of the most glaring areas where many of the meta-repositories for data systems did not align was in their use of metadata standards. Of the thirteen systems, only two used the same standard: DataCite and OpenAIRE (DataCite metadata standard). The remaining eleven systems all used a local standard—RIOXX, DDI Lite, panFMP, DDI, RIF-CS—or were not using a standard for various reasons. At one end of the spectrum was one system, EUDAT, that required only one metadata element for creating a record: a title. On the opposite end, DataBridge required the most metadata with over twenty-four elements from the DDI Lite standard. The most common elements found in meta-repository metadata schemes were title ( $n=10$ ) and author/contributor ( $n=6$ ).

## Discussion

This comparison revealed varying stages of development for each meta-repository. Many were just recently launched in the last five years, which means their systems may not have undergone many iterations to improve functionality or usability. Additionally, as many of these repositories overlap in content and mission, the ongoing availability is of concern. Federal and grant funds are often limited, thus many of these systems may be competing for the same funding streams.

The metadata issue is also incredibly significant. Without a common standard and element set it is doubtful that these systems will be fully interoperable. This issue is not limited to just meta-repositories; Moulaison, Dykas, and Galant found that roughly half of the twenty-three open-access repositories they surveyed were using the same metadata standard, Dublin Core.<sup>13</sup> The remaining used a combination of Qualified Dublin Core, MODS, and MARC. Additional-

ly, given the flexibility of many of these standards, the application of a standard varies both within and across systems. For example, dates can represent vastly different points given how a local repository makes use of the field. A date field can be interpreted as the date the asset was published online, the date the asset was created, and the date it was published in print. When a meta-repository pulls these two repositories together in the same system, the inconsistency is problematic. The answer to these issues might be to use computer systems to parse and normalize, or for the data repository community to come together and agree on a more rigid application of the metadata elements in a standard.

Additionally, while many of the repository aggregator missions were to support the accessibility and persistence of scholarship, few of the repository aggregators had facets that allowed users to limit to open-access materials. Although they claimed persistence as a priority, how this was facilitated was not evident across any of the meta-repositories for data. For example, none of the meta-repositories assigned DOIs or persistent identifiers to the metadata records they were aggregating, and few, if any, had curation treatment procedures in place for the metadata. Policies of how the meta-repository handles withdrawn records are not always evident.

Finally, many of these meta-repositories of data have come to act as *de facto* representatives of the smaller systems they aggregate or harvest from. Much like a traditional consortium, the meta-repositories can advocate for the interests of the other systems, recommend metadata standards, suggest best practices for metadata element values, and potentially create inventories of technical infrastructure for data repositories.

## Conclusion

Scholarly communication is in need of systems to pull together and link dispersed research objects. Just as Netflix revolutionized film discovery and rental, meta-repositories are needed to discover and highlight research from varying providers, make recommendations, show relationships between research and researchers, and make connections among the digital assets. The whole story of research, and the complete scholarly record, is more than just the final publication. It includes funder information, data sets, documentation, and code in many cases. The meta-repositories of data are one tool that seeks to address this issue. There exist many challenges to making these systems robust and operational enough to fit the scholarly communications need. Community involvement at the local level is integral to ensuring the success of these systems. Engagement with COAR, RDA, CODATA, or even the meta-repositories directly, ensures the ongoing viability of these useful systems.

## Notes

1. Peter J. Nurnberg, Richard Furuta, John J. Leggett, Catherine C. Marshall, and Frank M. Shipman III, "Digital Libraries: Issues and Architectures," in *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries. Austin, Texas, June 11-13, 1995* (Austin, TX: 1995), 147–53, <http://www.jcdl.org/archived-conf-sites/dl95/papers/nuernberg/nuernberg.html>.
2. Donald J. Waters, "What Are Digital Libraries?" *CLIR Issues*, no. 4 (July/August 1998), <http://www.clir.org/pubs/issues/issues04.html>.
3. A. O'Carroll, S. Collins, D. Gallagher, J. Tang, and S. Webb, *Caring for Digital Content: Mapping International Approaches* (Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy, 2013), <http://dri.ie/caring-for-digital-content-2013.pdf>.
4. Confederation of Open Access Repositories, "About COAR: Strategic Plan," accessed November 11, 2015, <https://www.coar-repositories.org/about/coar-ev/strategic-plan/>.
5. Confederation of Open Access Repositories, *COAR Strategy 2016–2018 and COAR Work Plan 2016–2017*, final version (Göttingen, Germany: COAR, November 5, 2015), <https://www.coar-repositories.org/files/COAR-Strategy-2016-2018-Final.pdf>.
6. Research Data Alliance homepage, accessed November 15, 2015, <http://rd-alliance.org>.
7. Research Data Alliance, "Repository Platforms for Research Data: Case Statement," accessed November 20, 2015, <https://rd-alliance.org/group/repository-platforms-research-data/case-statement/repository-platforms-research-data-case>.
8. Research Data Alliance, "Research Data Interoperability Case Statement," Google Doc, Accessed March 29, 2016, [https://docs.google.com/document/d/11XZBXIXSOE\\_d0n-1JsaYx2LhRwkwel5OQNHBaXKG-a\\_o/edit](https://docs.google.com/document/d/11XZBXIXSOE_d0n-1JsaYx2LhRwkwel5OQNHBaXKG-a_o/edit).
9. Committee on Data for Science and Technology (CODATA), "Our Mission," accessed November 1, 2015, <http://www.codata.org/about-codata/our-mission>.
10. World Data System, "About," accessed November 1, 2015, <https://www.icsu-wds.org/organization>.
11. Tyler Walters and Judy Ruttenberg, "SHared Access Research Ecosystem," *Educause Review* 49, no. 2 (2014), 56–57.
12. Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci, "The OpenAIRE Literature Broker Service for Institutional Repositories," *D-Lib Magazine* 21, no. 11/12 (November/December 2015), doi:10.1045/november2015-artini.
13. Heather Lea Moulaison, Felicity Dykas, and Kristen Gallant, "OpenDOAR Repositories and Metadata Practices," *D-Lib Magazine* 21, no. 3–4 (March/April 2015), doi:10.1045/march2015-moulaison.

## Bibliography

- Artini, Michele, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci. "The OpenAIRE Literature Broker Service for Institutional Repositories." *D-Lib Magazine* 21, no. 11/12 (November/December 2015). doi:10.1045/november2015-artini.

- Committee on Data for Science and Technology (CODATA). "Our Mission." Accessed November 1, 2015. <http://www.codata.org/about-codata/our-mission>.
- Confederation of Open Access Repositories. "About COAR: Strategic Plan." Accessed November 15, 2015. <https://www.coar-repositories.org/about/coar-ev/strategic-plan/>.
- . *COAR Strategy 2016–2018 and COAR Work Plan 2016–2017*. Final version. Göttingen, Germany: COAR, November 5, 2015. <https://www.coar-repositories.org/files/COAR-Strategy-2016-2018-Final.pdf>.
- Moulaison, Heather Lea, Felicity Dykas, and Kristen Gallant. "OpenDOAR Repositories and Metadata Practices." *D-Lib Magazine* 21, no. 3–4 (March/April 2015). doi:10.1045/march2015-moulaison.
- Nurnberg, Peter J., Richard Furuta, John J. Leggett, Catherine C. Marshall, and Frank M. Shipman III. "Digital Libraries: Issues and Architectures." In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries. Austin, Texas, June 11–13, 1995*, 147–53. Austin, Texas, 1995, <http://www.jcdl.org/archived-conf-sites/dl95/papers/nuernberg/nuernberg.html>.
- O'Carroll, A., S. Collins, D. Gallagher, J. Tang, and S. Webb. *Caring for Digital Content: Mapping International Approaches*, Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy, 2013. <http://dri.ie/caring-for-digital-content-2013.pdf>.
- Research Data Alliance. "Repository Platforms for Research Data: Case Statement." Accessed November 20, 2015. <https://rd-alliance.org/group/repository-platforms-research-data/case-statement/repository-platforms-research-data-case>.
- . "Research Data Interoperability Case Statement." Google Doc. Accessed March 29, 2016. [https://docs.google.com/document/d/11XZBXIxSOE\\_d0n1JsaYx2LhRwkwel5OQNHBaXKG-a\\_o/edit](https://docs.google.com/document/d/11XZBXIxSOE_d0n1JsaYx2LhRwkwel5OQNHBaXKG-a_o/edit).
- Research Data Alliance homepage. Accessed November 15, 2015. <http://rd-alliance.org>.
- Walters, Tyler, and Judy Ruttenberg. "SHared Access Research Ecosystem." *Educause Review* 49, no. 2 (2014): 56–57.
- Waters, Donald J. "What Are Digital Libraries?" *CLIR Issues*, no. 4 (July/August 1998). <http://www.clir.org/pubs/issues/issues04.html>.
- World Data System. "About." Accessed November 1, 2015. <https://www.icsu-wds.org/organization>.

