

## Repositories: Open Data Infrastructure

Repositories of various types are increasingly important in the preservation, discovery, access, and reuse of research data. This workshop is focused on fostering a discussion around the various interests that are in play in the development, use, and selection of specific repositories - both as sources for discovering and accessing data for reuse, but also for sharing and preservation of research project products.

## Open Data: Why?

Policies and drivers include federal and other funder requirements:

- White House Office of Science and Technology Policy memo, *Expanding Public Access to the Results of Federally Funded Research*<sup>1</sup>
- Gates Foundation Open Access Policy<sup>2</sup>

The list goes on - NSF, NIH, DoE, DoD...

Policy (sticks) aside, there is likewise research into the cultural, scientific, and social benefits of making data open and reusable (carrots). The Fair Guiding Principles<sup>3</sup> provide a nice overview as well as a pretty useful acronym of what *open* means regarding data:

Findable Accessible Interoperable Reusable

We note the NSF is currently funding projects around research and data re-use.

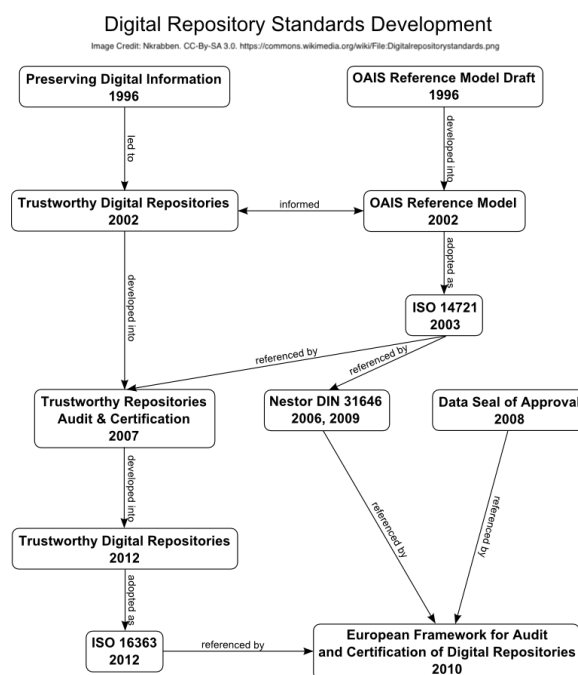
## Classes of Repositories

Repositories may be categorized in a number of different ways, depending upon needs:

Compliance with standards that define trustworthiness, measures of reliability, or other predefined characteristics, such as the

- Trusted Digital Repository (TDR) Checklist<sup>4</sup>, or its successor
- Trustworthy Digital Repositories - ISO 16363<sup>5</sup>/CCSDS 652.0-M-1<sup>6</sup>

### • Data Seal of Approval<sup>7</sup>



Types of content that those repositories focus on:

- **Disciplinary repositories** that specialize in content produced by specific research disciplines such as ICPSR<sup>8</sup> for social science data, the Archaeology Data Service<sup>9</sup> for archaeological data, and GenBank<sup>10</sup> for genetic sequence data.
- **General or Interdisciplinary repositories** that contain content that cross disciplinary boundaries, and may also contain multiple types of content including data, documents, and multimedia files.

Or, the organizations that host/manage the repositories - frequently as a designated repository for data products produced by or with funds from those organizations. For example:

- Data.gov<sup>11</sup> - the discovery portal for data generated and provided by US Federal Government agencies, and some states, municipalities, and universities

- NASA's Distributed Active Archive Centers (DAACS)<sup>12</sup> - the 12 distributed data centers supported by NASA that provide discovery and access to different thematic collections of Earth science data.
- UNM's digital repository<sup>13</sup> - the designated *institutional repository* for the University of New Mexico.

## Open Data: Where?

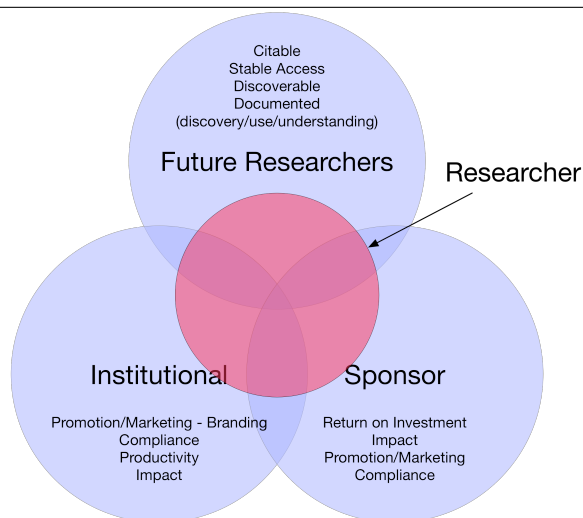
Requirements for repository services will vary from case to case, and there is no one-size-fits-all solution. Relevant factors include:

- Trusted repository status
- Capacity and commitment to preservation
- Provision of permanent identifiers (DOI)
- Format support
- Federation, discovery, and interoperability with other systems

Often these and other repository characteristics and policies are documented. Resources for identifying and assessing repositories include:

- PLOS ONE Data Availability policy (includes a list of recommended repositories<sup>14</sup>)
- Nature Scientific Data Recommended Data Repositories<sup>15</sup>
- Scientific Data recommended repositories<sup>16</sup>
- Registry of Research Data Repositories (re3data.org)<sup>17</sup>

## A Synthesis



## Further Reading

- Baker, Karen S., and Lynn Yarmey. "Data Stewardship: Environmental Data Curation and a Web-of-Repositories." *International Journal of Digital Curation* 4, no. 2 (2009): 12-27. <https://doi.org/10.2218/ijdc.v4i2.90>
- Beagrie, Neil, and John Houghton. "The Value and Impact of Data Sharing and Curation: A Synthesis of Three Recent Studies of UK Research Data Centres." London: JISC, 2014. [http://repository.jisc.ac.uk/5568/1/iDF308\\_-\\_Digital\\_Infrastructure\\_Directions\\_Report%2C\\_Jan14\\_04.pdf](http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_04.pdf)

Download this document: <https://unmrds.github.io/bb-discovery/bb-discovery.pdf>  
 Github Repository: <https://github.com/unmrds/bb-discovery>

## Notes

<sup>1</sup>White House Office of Science and Technology Policy (OTSP) (2013). *Expanding Public Access to the Results of Federally Funded Research*. <https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>

<sup>2</sup>Bill and Melinda Gates Foundation. *Open Access Policy*. <https://www.gatesfoundation.org/how-we-work/general-information/open-access-policy>

<sup>3</sup>Nature Scientific Data (2016). The FAIR Guiding Principles for scientific data management and stewardship. <https://www.nature.com/articles/sdata201618>

<sup>4</sup>RLG-NARA Task Force on Digital Repository and Certification (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Version 1, February 2007. Robin L. Dale and Bruce Ambacher, ed. The Center for Research Libraries & OCLC Online Computer Library Center, Inc. [http://www.crl.edu/sites/default/files/d6/attachments/pages/trac\\_0.pdf](http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf)

<sup>5</sup>International Organization for Standardization (ISO) (2012). *ISO 16363:2012 (CCSDS 652.0-R-1) Space data and information transfer systems – Audit and certification of trustworthy digital repositories*. <https://www.iso.org/standard/56510.html>

<sup>6</sup>The Consultative Committee for Space Data Systems (2011). *Audit and Certification of Trustworthy Digital Repositories Recommended Practice CCSDS 652.0-M-1*. <https://public.ccsds.org/pubs/652x0ml.pdf>

<sup>7</sup>Data Seal of Approval website (2018). <https://www.datasealofapproval.org/en/>

<sup>8</sup>Inter-University Consortium for Political and Social Research (ICPSR). <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>

<sup>9</sup>Archaeology Data Service (ADS). <http://archaeologydataservice.ac.uk/>

<sup>10</sup>GenBank. <http://www.ncbi.nlm.nih.gov/genbank/>

<sup>11</sup>Data.gov (2018). <https://catalog.data.gov/dataset>

<sup>12</sup>NASA Distributed Active Archive Centers (DAACS) (2018). <https://earthdata.nasa.gov/about/daacs>

<sup>13</sup>UNM's Institutional Repository (2018). <http://digitalrepository.unm.edu>

<sup>14</sup>PLOS ONE (2014). *Data Availability*. <http://journals.plos.org/>

[plosone/s/data-availability](https://plosone/s/data-availability)

<sup>15</sup>Nature Scientific Data (2018). *Recommended Data Repositories*. <https://www.nature.com/sdata/policies/repositories>

<sup>16</sup>Scientific Data (2018). *Scientific Data Recommended Repositories*. doi:10.6084/m9.figshare.1434640.v11

<sup>17</sup>Registry of Research Data Repositories (2018). <https://www.re3data.org/>

This work is licensed under a Creative Commons Attribution 4.0 International License - <http://creativecommons.org/licenses/by/4.0/>