MEASURING THE IMPACT OF DIGITAL REPOSITORIES

SUMMARY OF THE BIG DATA WORKSHOP

Prepared by the

BIG DATA INTERAGENCY WORKING GROUP

NETWORKING & INFORMATION TECHNOLOGY
RESEARCH & DEVELOPMENT SUBCOMMITTEE

COMMITTEE ON S&T ENTERPRISE

of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

JULY 2018

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at http://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the Big Data Interagency Working Group

Federal agency members of the Big Data Interagency Working Group (BD IWG) coordinate research and development (R&D) focused on improving the management and analysis of large-scale data to extract knowledge and insight from large, diverse, and disparate data sources, including mechanisms for data capture, curation, management, visualization, and access. The BD IWG works under the auspices of the Networking and Information Technology Research and Development (NITRD) Subcommittee of the NSTC Committee on Science and Technology Enterprise to identify current big data R&D activities across the Federal Government and to offer opportunities for coordination of R&D activities among agencies, academia, and the private sector. More information about the BD IWG is available at https://www.nitrd.gov/groups/bd.

Copyright Information

Key Takeaways

The Big Data (BD) Interagency Working Group (IWG) held a workshop, *Measuring the Impact of Digital Repositories*, on February 28 and March 1, 2017, in Arlington, VA. The aim of the workshop was to identify current assessment metrics, tools, and methodologies that are effective in measuring the impact[1] of digital data repositories, and to identify the assessment issues, obstacles, and tools that require additional research and development (R&D). This workshop brought together leaders from academic, journal, government, and international data repository funders, users, and developers to discuss these issues. Workshop participants identified five key improvements that would enhance the impact of digital repositories:

- A group with broad expertise and experience able to formulate and recommend best practices for data sharing and reuse.
- A data citation system that treats data as first-class objects comparable to publications in the research life cycle.
- Data repository certification that is understandable and usable across a broad range of repositories.
- New methods to assess economic impacts and opportunity costs when a repository is maintained or eliminated.
- A suite of strategies that repositories can use to achieve financial sustainability.

Background

There is a strong consensus in the scientific community that digital repositories are a critical component of the Nation's research infrastructure. The importance of this infrastructure is growing as repositories are evolving from storage spaces for basic raw data to facilities that support complex functionality, advanced data queries, and the use of specific analytic tools. In an environment where maintaining data repositories is expensive and resources are finite, efficiency demands developing new tools and methodologies to assess the impact of digital repositories; this development is necessary to be able to vouch for and communicate the authenticity, reliability, accessibility, and usability of the repositories. Systematic assessment approaches and well-understood metrics—both quantitative and qualitative—are needed.

The Big Data Interagency Working Group, which is co-chaired by the National Science Foundation and the National Institutes of Health, organized this workshop to bring together representatives from a broad range of disciplines and stakeholder groups to integrate existing knowledge and work toward an understanding of next steps. Workshop participants were asked to share their individual experiences and perspectives on how best to assess the impact of any given digital repository. Although the workshop was held prior to the release of the President's Management Agenda (PMA), the outcome of the workshop and the recommendations of the Big Data Interagency Working Group should be helpful in developing the integrated Data Strategy under the PMA.[2]

[1] "Impact" is defined here to mean both a measure of a data repository's success (or value) to the various user communities and as a driver of the innovation that broadly benefits the U.S. economy and its citizens.

[2] https://www.performance.gov/PMA/PMA.html.

The following is a synopsis of the discussions among the experts who attended the workshop.[3]

Event Focus

The workshop focused on two main areas: procedures and technologies to measure the impact of data repositories, and innovative strategies to improve their financial sustainability.

Key Topic 1: Developing new procedures and technologies for determining the impact of digital repositories

There are mature research communities today that have a culture of data sharing and reuse and that recognize the value of data repositories. The collection and dissemination of their best practices is an important first step to improve the overall data ecosystem, including in areas such as data reuse, collaboration across disciplines, and development of metrics for determining impact. Interaction among stakeholders from across the digital repository community (e.g., researchers, funders, digital preservationists, information professionals, and publishers) also is key to facilitating a productive exchange of ideas and solutions.

Current metrics vary widely, are not standardized, and their value largely depends on the stakeholder.[4] While metrics like web statistics and citation counts are in wide use and are easily countable, they fail to illustrate the full extent of repository activity. More qualitative assessments are needed for a fuller understanding of impact; such assessments could include documenting stories of user outcomes, the levels of trust that users give a repository's holdings, and whether a repository is used as a resource for training researchers.

Need for high-level leadership

There is a need to establish a body of both public and private sector data repository leaders who have both breadth and depth of expertise to share information on best practices, strategic investments, metrics, stewardship, and preservation policies. With regular meetings, transparent and accountable rules for membership, and discussions grounded in broadly accepted principles for data sharing and reuse,[5] this leadership body could:

- Develop metrics and methods of assessment that cover the full range of repositories from raw data repositories to complex databases that have advanced query and analytical tools. This is critically important for situations where databases are explicitly linked or where researchers need to integrate data from different sources.
- Develop key indicators and shared metrics for data repository sustainability.

[3] More workshop information is available at https://nitrd.gov/nitrdgroups/index.php?title=DigitalRepositories.

[4] York J., Gutmann, M., Berman, F., Will Today's Data Be Here Tomorrow? Measuring the Stewardship Gap, *Proceedings of the 2016 International Conference on Digital Preservation* (IPRES 16), October 2016, p. 102, https://services.phaidra.univie.ac.at/api/object/o:503172/diss/Content/get.

[5] The FAIR Data Principles (Findable, Accessible, Interoperable, and Reusable) are described at https://www.force11.org/group/fairgroup/fairprinciples.

- Promote data reuse by working with university research managers and researchers, especially early-career researchers, to understand the importance of both institutional and subject research repositories.

Need for better methods for data attribution

In a robust data repository ecosystem, data must be first-class objects that include a citation system. The value of data and the impact of data repositories are inextricably linked. Repositories can help change the culture of the research life cycle to one where data preservation is as important as publication. To accomplish this change, a data citation infrastructure is needed that partners with librarians, archivists, and journal publishers and provides methods for data retrieval and reuse. As a positive outcome, funders and peer-review groups will be able to easily assign credit to individuals and organizations and thus incentivize the deposit of useful data.

Need for better use and understanding of data repository certification

Certification methods are currently available to help judge the ability of a digital repository to preserve and provide access to its data. These range from cursory self-assessments to peer-review to full ISO (International Organization for Standardization) audit and certification. One example is the European Union's *European Framework for Audit and Certification of Digital Repositories*.[6] The Framework recognizes a sequence of three levels of certification—basic, extended, and formal—in order of increasing trustworthiness. While standards currently exist, knowledge of, and experience with these standards varies widely. Whereas some stakeholders express confidence in the ability of standards to ensure that a repository's holdings remain authentic, reliable, accessible, and usable on a continuing basis, others find the standards difficult to use and understand. Improving the applicability, usability, and knowledge of these standards can help all stakeholders better assess a repository's impact.

Need for new methods to assess the economic impact of data repositories

New methodologies are needed to assess the economic impact of data repositories, as well as the opportunity cost if a repository is eliminated. In January 2016, the European Molecular Biology Laboratory's European Bioinformatics Institute published a report on the value and impact of managing public life science data.[7] The report focused on traditional "market" indicators such as investment valuation, impact surveys, and return-on-investment valuations to understand the repository's value to their stakeholder community.

Key Topic 2: Improving the financial sustainability of data repositories

There are costs involved in both preserving data for reuse and in deleting it. Whether data repositories are publicly or privately funded, as costs rise, there is a growing need for innovative management strategies. Public-private partnerships and the formation of international consortia are two promising strategies to create sustainable business models for a wide variety of data communities.[8] For example,

[6] http://trusteddigitalrepository.eu/Welcome.html.

[7] Beagrie N., Houghton J., *The Value and Impact of the European Bioinformatics Institute*, 2016, https://beagrie.com/static/resource/EBI-impact-report.pdf.

[8] Anderson W., et al., *Towards Coordinated International Support of Core Data Resources for the Life Sciences*, November 18-19, 2016 (preprint), https://www.biorxiv.org/content/early/2017/04/27/110825.

an organization such as the National Academy of Sciences could study partnerships between funding agencies and private organizations such as Google, Amazon, and Facebook to adapt business-related data management infrastructures to support scientific research.

Conclusion

Leading digital data repositories currently have various assessment mechanisms in place, but they are inconsistent and limited in their ability to fully capture meaningful impacts. What they do provide is a starting point. The workshop identified five areas to further improve these metrics and build a set of best practices. Those areas are high-level leadership, better methods for data attribution, better use and understanding of digital repository certification, new methods to assess the economic impact of data repositories, and a variety of strategies to achieve financial sustainability.

The workshop also explored the need for innovative strategies and models for ensuring data repository sustainability, including the use of international consortia, public-private partnerships, and other resources to help build sustainable business models.