Big Data: Pioneering the Future of Federally Supported Data Repositories Workshop Report

Big Data
Interagency Working Group

Laura J. Biven and Amy Walton, Co-chairs

February 2022

Table of Contents

Executive Summary

On January 13–15, 2021, the Big Data Interagency Working Group (BD IWG) of the Networking and Information Technology Research and Development Program held a workshop on "Pioneering the Future of Federally Supported Data Repositories" to explore opportunities and challenges for the future of federally supported data repositories (FSDRs). FSDRs facilitate access to federally funded research data and play a pivotal role in enabling machine learning, artificial intelligence, and other data-driven science and discovery. FSDRs also play a critical role as building blocks for a future data ecosystem that emerged during the workshop.

The workshop brought together FSDR representatives, thought leaders in data science, and user communities to explore the vision for the future of FSDRs. The envisioned future is a data ecosystem optimized for data science that is characterized by innovative, user-driven approaches to combining data for insights. This data ecosystem requires capabilities for integrating data across FSDRs and other data sources that are agile, routine, and broadly enabled. After the workshop, the following findings were identified by the BD IWG members as enablers of this vision:

1. Strong partnerships between FSDRs and the science research community to continually enhance the potential for discovery and provide benefit to society.
2. Integration of FSDRs into an internationally recognized, national-scale data and computing ecosystem to broaden participation and enable faster, easier, and more equitable access to data-driven science.
3. Recognition and support for FSDRs as long-term infrastructure to provide continuity of service to the research community as well as a fertile environment for training future generations of scientists.
4. Participation by FSDRs in an interdisciplinary, national-scale community of repositories to enhance interoperability and achieve goals not otherwise obtainable from single data sources.
5. Foster a culture of continuous evaluation and improvement of FSDRs for greater scientific impact from data.

Note: Any mention in the text of commercial or academic partners in Federal R&D activities is for information only; it does not imply endorsement or recommendation by any U.S. Government agency.

Introduction

The Networking and Information Technology Research and Development (NITRD) Big Data (BD) Interagency Working Group (IWG) held a virtual workshop on "Pioneering the Future of Federally Supported Data Repositories" on January 13–15, 2021.[1] This workshop explored the opportunities and challenges for federally supported data repositories (FSDRs). Data-driven research, as well as artificial intelligence and machine learning (AI/ML), brings renewed focus to the data in these repositories. Workshop participants explored how these repositories should adapt to the emerging and evolving needs and requirements of data-intensive research, what can be done to prepare for this future, and what is needed to build and strengthen the FSDR community. The workshop engaged FSDR representatives, thought leaders in data science, and user communities to address these topics.

Federal initiatives on open science and open data, along with NITRD[2] workshops and strategic plans for AI/ML[3] and the Future Advanced Computing Ecosystem (FACE),[4] recognize the importance of access to data, be it open or restricted. The interest in open data and data repositories is global, with large-scale efforts in data science infrastructure and data governance underway in Europe, Australia, China, and Asia. FSDRs constitute the building blocks of a findable, accessible, interoperable, and reusable (FAIR) data ecosystem and play a key role in data access and use. The FSDRs are united by common interests and challenges such as the ever-increasing need for data interoperability and reusability across repositories; a rapidly changing and evolving landscape with powerful new storage and computing platforms and capabilities; growing expectations from users, funders, and publishers; and increasing expectations regarding privacy, security, integrity, ethics, bias, and equity.

The workshop included close to 200 participants representing expertise across a wide range of scientific disciplines from academia, national laboratories, Federal agencies, industry, and nonprofit organizations. This report summarizes the salient aspects of pioneering the future of FSDRs and provides the key takeaways from the workshop.

Key Findings

During the workshop, the participants discussed the need to bring together data from multiple FSDRs to impact critical research and societal challenges. Some examples include the need for integrating massive amounts of microbiome data from multiple sources with spatial and theoretical frameworks to better understand the relationships between individual human microbiome states and broader concepts of health and disease; providing tools for distributed analyses of large-scale climate data together with information about population and social data; and integrating real-time data streams to support improved, evidence-based decision making in battle. Currently, analyzing data from multiple FSDRs is a burdensome task. An ambitious goal of the envisioned future for FSDRs is a data ecosystem optimized for data science that is characterized by innovative, user-driven approaches to combining data for insights. The future data ecosystem therefore necessitates capabilities for combining data across FSDRs and other data sources that are agile, routine, and broadly enabled, thus presenting challenges that require technical, social, and practical solutions. The following findings were identified during the workshop as enablers of this vision.

[1] FSDR Workshop Program, Jan 13–15, 2021, https://www.nitrd.gov/nitrdgroups/images/4/43/Federally-Supported-Data-Repositories-Agenda.pdf

[2] https://www.nitrd.gov/

[3] https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

[4] https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf

Partnership with the Science Research Community

Strong partnerships between FSDRs and the science research community to continually enhance the potential for discovery and provide benefit to society.

Workshop participants recognized the need for FSDRs to engage proactively with the community to continuously identify strengths and gaps and anticipate the needs of the research community, including use cases in AI/ML. Currently, researchers spend considerable time and effort understanding aspects of data and overcoming barriers that have a substantial impact on their analyses; for example, identifying sources of noise, uncertainties in data and data labels, and imbalances or biases in the data. FSDRs can help reduce these barriers by working with the research community to develop innovative approaches to documentation, tool development, and engagement and support. Codeathons/Hackathons or other events for coders to work collaboratively on software development were suggested as a way for FSDRs to address this need.

In addition, participants emphasized the need for FSDRs to be proactive in developing new capabilities such as automated curation, quality assessments, and bias measures. Workshop participants discussed support for scientific use case development, cataloging, and sharing, as well as for partnering with researchers to exploit heterogeneous data across repositories, as an additional way that FSDRs can collaborate in scientific advancements. FSDRs could assist users in the creation of a variety of useful data services, such as generating synthetic data to capture as much saliency as possible from real data, developing modeling services that are able to gain value from inconsistent data without lowering overall performance, and accelerating development of AI/ML and other methods that use small amounts of data. FSDRs could also help promote the ethical use of data by remediating imbalances and biases when possible, providing information and guidelines to users, and engaging stakeholders to consider the ethical questions in exposing data to new analysis techniques and technologies.

Participants saw an opportunity for FSDRs to help set new intellectual directions by assisting researchers with tackling data-driven questions:

- How much data do I need for a given study?
- Are two different datasets consistent with each other and equally good to answer a particular question?
- What minimum new data do I need to add for optimal benefit?
- Is a given dataset "balanced"/unbiased for a particular use but not for another?
- Can I use datasets from different studies to increase effective sample size?
- What is the scale of data needed for effectively answering a particular question?

Finally, participants discussed how FSDRs could also help researchers make their data FAIR and AI/ML ready. Engagements with the community could help with the development of requirements, standards, and guidelines for FAIR and AI/ML readiness as well as promote the use of these best practices.

National-Scale Data and Computing Ecosystem

Integration of FSDRs into an internationally recognized, national-scale data and computing ecosystem to broaden participation and enable faster, easier, and more equitable access to data-driven science.

During the workshop, it was noted that realizing data's full potential requires a more seamless and equitable ecosystem so that elements of data science workflows are not siloed in independently provisioned environments and that access to data science capabilities is not limited by a lack of local resources. Data science, which is characterized by innovative, user-driven approaches to combining data to derive new insights, emerged as an organizing principle for the future ecosystem.

The workshop participants noted the opportunities to accelerate data science with a national-scale ecosystem with which users can easily bring data to cutting-edge computing platforms. Currently, many workflows involve time-consuming and/or manual access, preparation, and movement of data. There is an opportunity to build cyber infrastructure that makes data science faster and easier. Another key opportunity identified was broader and more equitable access to data, computing, and tools to meet the growing interest in data science. Solutions to advance these aims need to include data repositories, both in the cloud and on-premises, as well as federated or distributed approaches to data science computation.

The emerged vision for the ecosystem includes the FAIR data principles as well as interoperable, scalable, and portable tools and platforms to rapidly facilitate new use cases and workflows that can better integrate, combine, or analyze data. Other key components and attributes of the future ecosystem identify core data services such as ontologies, interoperability standards, persistent unique identifiers, and access and search capabilities that are sustained and available; easily available computing and storage cyber infrastructure co-located with data; a research culture that values, prioritizes, and rewards reuse of data and other research products; and data management practices and expertise.

Public-private partnerships and collaborations, including those with foreign partners, were also seen as essential. Commercial vendors and high-tech companies are significant providers and innovators in cyber infrastructure, data management, and data analysis. Internationally, efforts in data interoperability are underway with significant community involvement.

Data services (e.g., metadata standards, knowledge graphs, and visualization platforms) can dramatically improve and democratize the usability of data, as has been shown through the success of geographic information systems for geospatial data.

Long-Term Infrastructure

Recognition and support for FSDRs as long-term infrastructure to provide continuity of service to the research community as well as a fertile environment for training future generations of scientists.

Workshop participants discussed the vital importance of long-term support for FSDRs. For example, a critical threat to the community's ability to maintain and derive value from data is the lack of trained experts in data management and data engineering. The participants noted that long-term support for FSDRs encourages the development of a trained workforce that helps to develop, maintain, and operate these resources by providing job opportunities and, importantly, career paths in FSDR management. It will encourage expert repository managers and stewards not only to address operational needs but also to conduct research in advanced repository management areas.

Participants noted that sustained support would also allow FSDRs to address long-term operational and stewardship goals, particularly those that incur risk. This support could provide services and technology components that are more efficient, reusable, and cost-effective. Some notable examples shared by the participants included opportunities to develop AI-based tools to assist with repository management, data cleansing and curation, and anticipation and meeting of user needs; continued research in data security, federated search, discovery, and analysis; and automated approaches for metadata capture. These services and technology components enable dynamic and rapid reconfiguration of infrastructure to accommodate multimodal and other data types. They also provide new and smarter methods in repository management for capturing, preserving, and tracking the provenance of digital assets, including transient or dynamic data, which are not adequately addressed in today’s repositories. Privacy and ethics were other areas in which participants noted that long-term support would encourage innovation and adoption of best practices.

Consistent data security and privacy approaches adopted across Federal repositories were also identified as a critical need not only to implement effective governance and avoid security breaches but also to allow for opportunities in advancing security research and development (R&D) as a research domain. The participants discussed the challenge of addressing ethical, legal, and social implications, given the increased security risks and new privacy laws. The need to balance open science while complying with privacy laws appears to be a critical gap, as is the need for computational R&D to continue to allow for computational approaches in unbiased, data-driven research and innovation.

Sustainability of FSDRs beyond or in addition to Federal agency support was a topic of significant discussion, including the following ideas:

- Repositories to develop alternative business models to sustain operations.
- Various innovative approaches to address operational efficiency by making services and technology components more efficient, reusable, and cost-effective.
- Sustainability of staffing from the lack of well-trained professionals and resources needed to keep up with the increasing efforts to maintain secure access and ensure high-integrity data sources.

Interdisciplinary, National-Scale Community

Participation by FSDRs in an interdisciplinary, national-scale community of repositories to enhance interoperability and achieve goals not otherwise obtainable from single data sources.

During this discussion, a vision emerged of national-scale or global data service, solutions for curation, reduction of redundant efforts across the FSDRs, and interaction with common stakeholders. Following are some of the potential next steps identified to foster this national-scale community:

- Expanding existing (and planned) **community-level data services** such as search, ontologies, and metadata services can facilitate and encourage cross-disciplinary collaborations.
- **Envisioning the future of data publication** could lead to new paradigms, potentially similar to current software publishing. The function of peer review may be better served by community codeathons, while the concept of Version of Record should adapt to the dynamic nature of data.
- Developing more efficient and effective **publication tracking** resources could reduce redundant FSDR efforts.
- Creating better solutions for **sustainable data curation** can play a critical role in contributing expert input while helping to alleviate the prohibitive costs if these efforts were exclusively staffed by the FSDR.
- Better **recognizing and fostering roles of the various stakeholders** in the community could optimize benefits, such as those of libraries/archives that are developing digital information systems that could be better leveraged by research repositories.
- **Establishing partnerships with one or more research repositories**, both domain specific and general, can help foster communities of practice.

Based on the discussions, there appears to be significant opportunity for community coordination that is multidisciplinary and multi-sector in nature with engagement from academia, industry/for-profit, nonprofit, and Federal/governmental organizations at an international scale. All sessions suggested that FSDRs should consider taking a leading role to foster communities of practice that include all the stakeholders, a model recently proposed by the Research Data Alliance.[5] Outside of workshops, promoting ad hoc visits between FSDRs could help identify potential opportunities for collaboration.

[5] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

Evaluation and Improvement

Foster a culture of continuous evaluation and improvement of FSDRs for greater scientific impact from data.

Workshop participants noted the importance of measuring the impact of data. There was a clear interest in developing meaningful and consistent impact measures and other indicators that address data's current and potential value. One example is to mirror the FAIR principles with "FAIRed" impact measures on how data have been found, accessed, interoperated, and reused. Workshop participants also discussed the need for clear and consistent reporting on usage and utility; for example, by developing data use scorecards. Leveraging community-established forums to develop key performance indicators (KPIs) that measure the repository's impact and usage, as well as the data's scientific impact, can help determine long-term usefulness and value.

It was also recognized that the value of data may need to be evaluated over time and in combination with other data. Measuring scientific impact is not trivial, and measuring true utility requires new approaches, such as KPIs. Participants mentioned an international effort to standardize data usage metrics,[6] which is gaining popularity as a framework for a community-driven approach to innovation based on KPIs.

Conclusion

This workshop convened experts and representatives of FSDRs to discuss the future of the FSDR assets in light of recent changes in analysis tools and needs, policy expectations, and their potential to underpin a data science ecosystem that enables innovation and data-driven research. The findings from the workshop, as determined by the BD IWG, suggest the need for a stronger, more innovative role for FSDRs, and a dynamic data and computing ecosystem in which new use cases that combine data and tools in unique ways are enabled quickly and at scale. The workshop discussed an ambitious, future vision for FSDRs, and possible future directions to advance this goal. A national-scale data ecosystem could be an important component for responding to national priorities as well as for advancing innovative, data-driven science.

[6] https://makedatacount.org/

List of Abbreviations and Acronyms

Item	Spell-out
AI/ML	artificial intelligence and machine learning
BD	Big Data
DoD	Department of Defense
DOE	Department of Energy
FACE	Future Advanced Computing Ecosystem
FAIR	findable, accessible, interoperable, and reusable
FSDR	federally supported data repository
IWG	Interagency Working Group
KPI	key performance indicator
NCO	National Coordination Office
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development
NSF	National Science Foundation
R&D	research and development

About the Authors

The NITRD Program is the Nation’s primary source of federally funded work on pioneering information technologies in computing, networking, and software. The NITRD Subcommittee of the National Science and Technology Council’s Committee on Science and Technology Enterprise guides the multiagency NITRD Program in its work to provide the R&D foundations for ensuring continued U.S. technological leadership and that meets the Nation's advanced IT needs. The National Coordination Office (NCO) supports the NITRD Subcommittee and the IWGs and teams that report to it. The NITRD Subcommittee’s Co-Chairs are Kamie Roberts, NCO Director, and Margaret Martonosi, Assistant Director of the NSF Directorate for Computer and Information Science and Engineering. More information about NITRD is available online at https://www.nitrd.gov/.

The NITRD Program’s BD IWG coordinates Federal R&D that advances the ecosystem needed for extraction of knowledge and insights from data. More information about the BD IWG is available online at https://www.nitrd.gov/coordination-areas/big-data/.

Acknowledgments

The NITRD BD IWG gratefully acknowledges the workshop planning committee members Chaitan Baru (NSF), Laura Biven (NIH), Rajeev Agrawal (DoD), Ishwar Chandramouliswaran (NIH), Wo Chang (NIST), Frances Carter-Johnson (NSF), Ji Hyun Lee (NITRD NCO), and Jordan Thomas (DOE), who helped plan the workshop and to write and review this report; external advisors Debra Agarwal (Lawrence Berkeley National Laboratory), Philip Bourne (University of Virginia), and Julia Lane (New York University) for their valuable insights that helped to frame the workshop topics; and all the workshop participants for their contributions to the workshop discussions.