FAIR Data for Large Research Facilities

Don Brower
Center for Research Computing University of Notre Dame Notre Dame, Indiana USA dbrower@nd.edu

David Butcher
CIMAR, ICR
National High Magnetic Field Laboratory Tallahassee, Florida USA dbutcher@magnet.fsu.edu

Angela Murillo
School of Informatics and Computing
Indiana University-Purdue University Indianapolis, Indianapolis, Indiana USA apmurill@iu.edu

ABSTRACT

This workshop will bring together data managers, repository managers, administrators, and others who are responsible for, or interested in research data management at large research facilities. These facilities have unique issues due to a variety of factors, such as an extreme data volume, variety, and velocity. The workshop aims to provide cross-pollination between facilities that have similar desires to realize the FAIR principles. The organizers of this workshop are members of the NSF CI Compass FAIR Data Working Group, and the outcomes from these discussions will become a white paper and topics for future CI Compass webinars.

CCS CONCEPTS

• Information systems~Information systems applications~Digital libraries and archives •Applied computing •Social and professional topics~Computing / technology policy~Government technology policy

KEYWORDS

Digital Libraries; Cyberinfrastructure; Data Management; Persistent Identifiers; FAIR data

Introduction

Large research facilities have many unique data management challenges. Some challenges come from the significant and complex cyberinfrastructure required to collect, process, and store diverse datasets. Other challenges come from the difficulty of making the data findable, sharable and reusable to other researchers. The FAIR principles [1] are a codification of desirable properties for humans and their machine agents to interact with these data cataloging and discovery systems. The challenge with FAIR is that the principles are abstract and do not provide a concrete plan for how to meet them.

As science evolves, the need to make data available to others is increasing. Machine learning and artificial intelligence (AI) models require high-quality, trustworthy data for training. The value of these datasets comes from their size, both in completeness, and in high-dimensionality [2, 4]. These AI models are recognized as essential for the development of the next generation of data analysis methods and fostering innovation across the fields of science and engineering.

Lastly, recent changes in governmental mandates [3] have expanded the scope of data that needs to be

made available to all data collected for publicly funded research, whether related to a publication or not. All new federal funding will require a plan for making all data (with a few exceptions) available for reuse.

The CI Compass project (NSF award 2127548) is funded to engage with NSF Major Facilities to enable knowledge-sharing across MFs and the Cyberinfrastructure community. The CI Compass FAIR Working Group has been holding monthly meetings on data management and data infrastructure at MFs.

Topics

This half-day workshop will discuss issues related to data management that are of interest to large research facilities. This includes the following list of topics:

- The impact of the fall 2022 OSTP Nelson Memo [3] on NSF major facilities and other large research facilities.
- Persistent identifiers and how they are integrated into workflows. DOIs on datasets and software; ORCID; Instrument identifiers.
- Strategies of identifying, releasing, and versioning datasets and time-series data.
- Ways of tracking metadata for provenance and computational steps in data processing.
- Impacts of cybersecurity on data management and FAIR.
- Data lifecycle planning as an explicit element in facility cyberinfrastructure planning.
- Pragmatic ways of approaching FAIR-ness for large facilities.

The workshop will consist of a mixture of presentations and panel discussions. Through these discussions we hope to identify best practices, enlarge the community participating in these discussions, and solicit concerns. A tentative schedule follows:

1. Welcome and Introduction
2. Nelson Memo and Federal Data Management Guidelines
3. Outcomes of the CI Compass FAIR WG survey on FAIR strategy and implementations
4. Persistent Identifiers: implementation, pain points and best practices
5. Cybersecurity and FAIR

The workshop will be an in-person event. However, we may have virtual presenters based on schedule and travel constraints.

Audience

This workshop will be useful to anyone with responsibilities or an interest in data management or data infrastructure planning and operations at a large research facility. It will also be of interest to those with an interest in research data management, persistent identifiers, and cyberinfrastructure.

Outcomes

At a broad level we hope to foster a community of practice for data management at large research facilities. More specifically, participants will learn about the new US federal data management requirements and strategies for meeting them.

Topics, problems, and solutions identified through this workshop will be used by the NSF-funded CI Compass FAIR Data Working Group, to create webinars, technical reports and whitepapers to inform the wider community.

Speakers

- David Butcher (National High Magnetic Field Laboratory) is a scientist at the MagLab and works with the implementation of FAIR data management practices across the entire facility.
- Don Brower (Notre Dame) has experience with research data repositories, data preservation, and FAIR data.
- Brian Minihan (ORCID) is the Asia-Pacific Engagement Manager for ORCID. He has experience with persistent identifiers, data management policy, and digital libraries.
- Bruce Berriman (Caltech/JPL) has experience with FAIR data in astronomy through his involvement in the IVOA and his broader experience in data management.
- Angela Murillo (IUPUI) research is on data curation, scientific data management, and data cyberinfrastructure. Her experience covers industry and government agencies.

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

REFERENCES

[1] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci Data, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

[2] L. Christopherson, A. Mandal, E. Scott, and I. Baldin, "Toward a Data Lifecycle Model for NSF Large Facilities," in Practice and Experience in Advanced Research Computing, Portland OR USA, Jul. 2020, pp. 168-175. doi: 10.1145/3311790.3396636.

[3] A. Nelson, "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research." Office of Science and Technology Policy (OSTP), Aug. 25, 2022. [Online]. Available: https://www.whitehouse.gov/wp- content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

[4] NAIRR Task Force, "Envisioning a National Artificial Intelligence Research Resource (NAIRR): Preliminary Findings and Recommendations," May 2022. [Online]. Available: https://www.ai.gov/wp-content/uploads/2022/05/NAIRR- TF-Interim-Report-2022.pdf