A Survey of Scientific Data Needs at Oak Ridge National Laboratory

Suhas Somnath and Olga Kuchar

National Center for Computational Sciences and Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory

Introduction

- Inability to effectively manage and harness the scientific data leads to lost opportunities for scientific breakthroughs and significant losses in research productivity.
- 45 scientists and staff at Oak Ridge National Laboratory (ORNL) surveyed to identify needs in scientific data infrastructure and governance.
- Findings point to urgent need for comprehensive data infrastructure that spans across DOE complex and data governance.

Survey

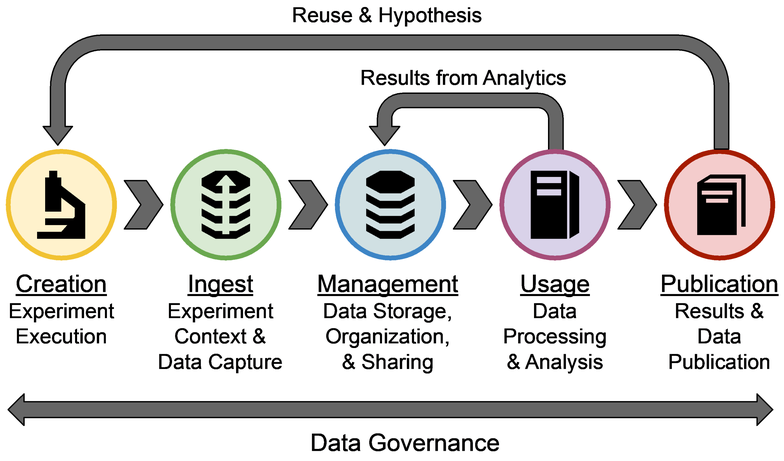

Figure 1: Lifecycle of a scientific dataset from creation to publication or destruction along with overarching data governance

- 45 researchers spread over all directorates in ORNL, working on a broad array of unclassified topics ranging from biofuels, to climate science were interviewed.
- Interviewees expressed data needs and challenges first to broad, open-ended questions and then to questions structured along the lifecycle of datasets and data governance, as shown in figure 1.
- Responses were manually synthesized, generalized (domain specificity), and collated.

Results

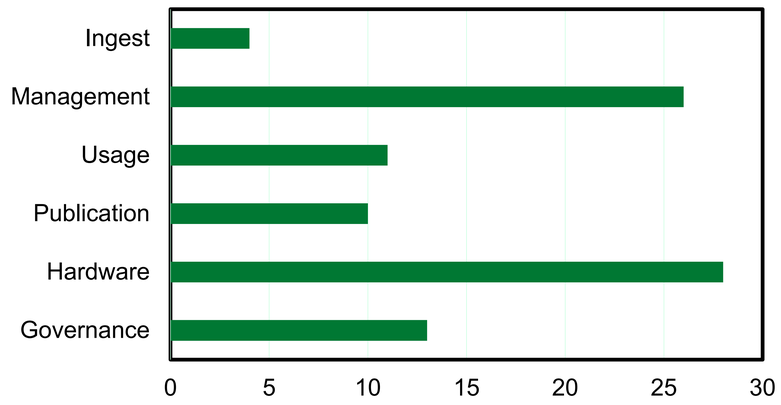

Figure 2: Popularity of data challenges faced by researchers.

Data Ingest

- 4 research groups expressed the need for a tool that:
  - Can be used by bench-scientists, at off-network instruments and even supercomputers.
  - Users can drag-and-drop data files from instruments.
  - Captures metadata / context regarding experiments.
  - Uploads this information to storage solutions, databases, information management systems, etc.

Data Management

- In this second largest category, researchers expressed an urgent need for an application that:
  - Provides comprehensive data searching, sharing, movement, organization, and collaboration.
  - Has user-friendly web and programming interfaces.
  - Provides user-friendly upload and download of data.
  - Complies with security clearances and stipulations
  - Tracks provenance, physical objects like samples and inventory in addition to data
  - Intelligently moves and caches data where necessary
  - Supports Findable, Accessible, Interoperable, and Reusable (FAIR) data principles.

Results (continued)

Data Usage

- **Workflows** – Researchers need tools that can help them easily craft data pipelines that span multiple machines (instruments, compute resources, etc.)
- **Analytics** – Researchers need lab-wide deployment of popular analytics services such as JuyperHub and Shiny R servers that can access and process data
  - There is a need for GPU accelerated analytics platforms for machine learning and deep learning

Data Publication and Cataloging

- Most researchers felt that the barrier to data publication could be lowered via:
  - User-friendly website and APIs to publish data
  - Funding and support to clean data for publication
  - Domain-specific data catalogs

Hardware

- The most popular data-need category was hardware:
  - **Storage**:
    - Large, reliable, and resilient storage solution.
    - A federation of existing data repositories
  - **Networking**:
    - Infrastructure to maximize throughput of streaming workflows processing large volumes of data
    - Network to move data from air-gapped instruments

Data Governance and other

- Besides infrastructure, researchers need:
  - Guidance in developing, implementing, and complying with data management plans for facilities and projects
  - Education on using tools and services

Acknowledgement:

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.