Scientific Data Management – an Under-recognized Grand Challenge

Suhas Somnath, Dale Stansberry, Joshua Brown, and Olga Kuchar

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

Introduction

- The explosion in scientific data volumes and generation rates and increasingly global and multidisciplinary nature of scientific research necessitate comprehensive scientific data management (SDM) capabilities.
- DM is crucial to maintain high research productivity and for scientific reproducibility due to the increasingly important role that data plays in scientific discovery.
- However, the adoption of DM tools like iRODS, Rucio, and DataFed has been slower than expected.
- We explain why DM is a grand challenge as shown in figure 1 by overviewing technical and cultural challenges that impede the adoption of scientific DM tools and propose solutions to overcome challenges.

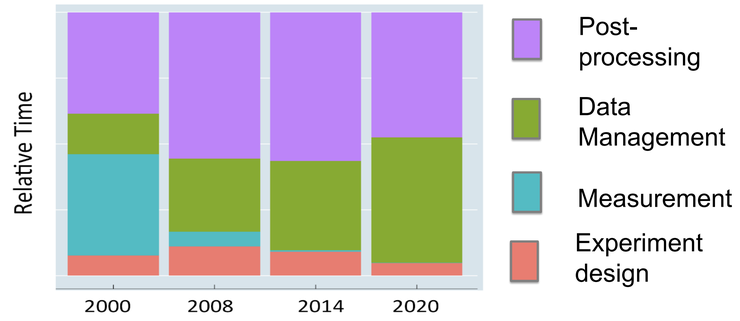

Figure 1: Growing importance of data management in synchrotron facilities.

Source – https://rawgit.com/4Quant/SRI2015/master/SRIPres.html

Cultural Challenges

1. Incentives

- Publications still the primary and only driver for success.
- Ad-hoc practices persist since academia and sponsors provide no reward for good SDM or penalty for bad DM

2. Learning Curve and Adoption Cost

- Comprehensive SDM often necessitates unlearning ad-hoc practices and learning SDM best practices and tools
- Researchers unable / unwilling to spend time learning
- Researchers don't have time to develop domain specific metadata adapters to integrate with SDM tools

3. Faith to Adopt DM Tools

Researchers hesitant to adopt SDM tools due to fear of:

- inadequate long-term support for SDM tool
- data getting locked within SDM tool and perceived effort necessary to extract data.
- community not adopting SDM tool to collaborate with, resulting in the classic "chicken-and-egg" problem.

Technical Challenges

1. Metadata

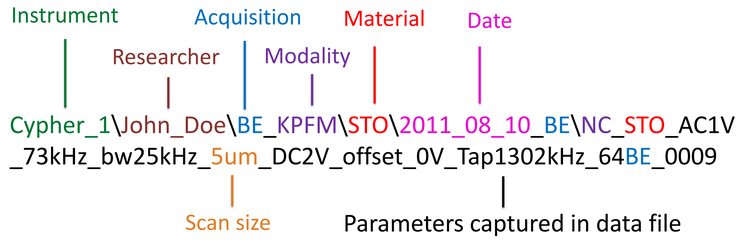

Figure 2: A real example of the current ad-hoc practice of embedding scientific metadata into file paths in order to be able to use file systems to search for data

- Researchers search for data using metadata as shown in figure 2. SDM tools rendered ineffective by unavailability or incompleteness of metadata.
- **Extraction**: Scripts and tools necessary to extract metadata from the thousands of file formats
  - Extractors may be available, incomplete, inaccurate, use different software stacks or simply unavailable.
  - Extractor scripts scattered and not cohesive.
- **Capture**: Inadequate tools to capture metadata from a process in workflow or in physical lab notebooks.
- **Schema**: Data challenging to search for if researchers use ad-hoc or unique schemas to represent metadata.

2. Generality vs Domain Specificity

- SDM tools need to support all scientific domains for reusability / scalability and multidisciplinary endeavors
- However, researchers desire domain-specific data visualization, search, analytics, ingest capabilities

3. Security

- Challenging for SDM tools to implement potentially abstract security guidelines for sensitive data
- Not scalable for SDM tools to integrate with each institution's unique databases to track security clearance of individuals

Proposed Solutions

DOE has the unique capability to solve these grand challenges in SDM through sheer scale:

1. Coordinated Cross-DOE Development Efforts

Coordinated efforts spanning DOE complex to develop:

- a robust, modular and reusable SDM tool
- comprehensive library for metadata extraction and capture a data publication and cataloging tool
- community standards metadata schemas

2. Long-term Commitment

- Long-term commitment to support development, maintenance, and deployment of tools and standards well past conventional 3-year projects

3. Incentives for Researchers

- Incentivize SDM best practices by rewarding the creation and use of existing published data similar to journal / conference publications
- Require that publicly funded research entail publication of data and be repeatable / reproducible

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.