Metadata and Reproducibility: A Case Study of Gravitational Wave Research Data Management

Jian Qin  Brian Dobreski  Duncan Brown
School of Information Studies  Department of Physics
Syracuse University  Syracuse University
GW research: Big science, big data, and big problem

Big science

Big data

Big teams

Complex workflows with entangled dependencies

Lack of metadata standards suitable for specialized domain

Data artefacts are not linked for easy tracking

Documentation of data is time consuming and inconsistent
Project goal

- Funded by NSF in 2014
- Goal:
  - To make the development, operation, and management of scientific workflows simpler, and the sharing and verification of their results more straightforward through metadata-aware infrastructure services.
- One of tasks is to develop a metadata model for representing workflows and associated artefacts.
Goal of the interview study

To understand and identify requirements for data management and discovery in GW research and develop a metadata model for representing the concepts and relationships.
Questions to be addressed

• What metadata is needed for GW researchers to track components in an analysis/search project?

• What metadata functions do researchers consider the most important in supporting GW research?

• How should the metadata model represent the needs for tracking, verification, and reproduction of GW science?
Gravitational Waves research: Computationally intensive science

Metadata

- Represents and describes
- Discovers, selects, and obtains
- Facilitates sharing, uses, and reuses

Reproducibility

Requires data and code to be
- in good quality
- discoverable
- verifiable
- trackable

Provenance support
Gathering user requirements

• Understand scientists’ data practice and priorities
  – Producing verifiable computing results requires subtle adjustment to work habits
  – Any adjustment must be based on a thorough understanding of workflows, data flows, and priorities in research lifecycle
Collecting initial information

- Initial information were collected and generalized from:
  - LIGO Scientific Collaboration (LSC) – Virgo website
  - Wiki pages of LSC working groups
  - Sample configuration files and related outputs
  - Metadata at the LSC Document Center
  - Documentations for databases
  - Meetings with scientists
The case of Gravitational Waves research data
Qualitative data collection

• Interview protocol was approved by Syracuse University Institutional Review Board

• Interview was designed to:
  – Understand research data and analysis needs and habits of GW scientists;
  – Learn about the requirements for discovering, tracking, and documenting, and archiving data and analysis products;
  – Define types of entities and relationships among the entities as well as current systems used for identification management of these entities; and
  – Collect vocabularies (category lists, terminology for instruments, parameters, data status and operations, analysis methods and techniques, etc.).
Interview questions

• Part 1: profile
  – Position in LSC
  – Time working in this field
  – Role in research group
  – Motivation for keeping documentation

• Part 2: data practice
  – Describing process of analysis
  – Things to do around starting an analysis
  – Most useful search options in data and/or workflow discovery
  – Effort required to reproduce an analysis
  – Information required to trust someone’s workflows
  – Any extra documentation
  – The most time-consuming part of managing workflows
  – Top three important features of a workflow management system
  – The ideal data management system
Participants’ profiles

• Eight participants
  – Selected based on the role and specialty to cast a wide coverage of different areas of GW research
  – Specialized in particle physics, gravitational physics
  – All in academic institutions: 4 in U.S. and 4 in Germany

  – Two full professors, five post docs, one graduate student
  – Years with LSC range from 4-15
  – Roles played in LSC groups:
    • Group chair
    • Search specialist
    • Code and pipeline developer
    • Data analyst
    • Detector characterization specialist

  – All participants are male
Findings

Motivation for keeping good documentation

Metadata that is against the grain

Reproducibility
Findings: Motivation for keeping good documentation

- Trust
  Verifiable code or pipelines is the basis for trust and reuse

- Accountability
  - “Exactly how you did what you did”
  - Accountability to funding agency

- Continuity
  Documentation supports continuity among dynamic community memberships
Findings: Reproducibility

• Code dependencies
• Code versions
• Documentation
• Content reproducibility vs. code reproducibility
• Verification
Findings: Metadata that is against the grain

- Data calibration and preparation
- Pipeline generated
- Run completion
Discussion

• Metadata for a complex series of interdependent executables and intermediate computing results is not always captured and organized

• To address the challenges, readily available provenance metadata for executables (creator, version, location, etc.)

• Important to GW research data management:
  – Author metadata (who originated the workflow, configure file, etc.)
  – Metadata for workflows inputs and data products
  – Metadata for relationships between authors and workflow inputs and data products (tracking who created what)
Summary

• Provenance data needs are high during the entirety of the research lifecycle

• There are two types of reproducibility in GW research: reproducing code and reproducing content and metadata facilitates both scenarios

• Identifying all inputs, outputs, and processes, their provenances, and the relationships between them is challenging but vital for this reproducibility

• As such, a metadata model specifically for GW research within the LIGO community is justified
Preliminary metadata model based on user requirement study
New discovery, new challenge: rethinking the metadata model

The data and code artefacts surrounding the Detection created a richer and new sources for the metadata model. New round of metadata model development work will be our next phase of work.
This project is funded by U.S. NSF grant #ACI-1443047

Thank you!

Questions?