IJDC | *Peer-Reviewed Paper*

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

Jian Qin
Syracuse University

Brian Dobreski
Syracuse University

Duncan A. Brown
Syracuse University

Abstract

The complexity of computationally-intensive scientific research poses great challenges for both research data management and research reproducibility. What metadata needs to be captured for tracking, reproducing, and reusing computational results is the starting point in developing metadata models to fulfil these functions of data management. This paper reports the findings from interviews with gravitational wave (GW) researchers, which were designed to gather user requirements to develop a metadata model. Motivations for keeping documentation of data and analysis results include trust, accountability and continuity of work. Research reproducibility relies on metadata that represents code dependencies and versions and has good documentation for verification. Metadata specific to GW data, workflows and outputs tend to differ from those currently available in metadata standards. The paper also discusses the challenges in representing code dependencies and workflows.

Received 20 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Jian Qin, School of Information Studies, Syracuse University, 311 Hinds Hall, Syracuse, NY 13244, USA. Email: jqin@syr.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Data-driven, computationally-intensive scientific research often involves complex workflows that contain entangled dependencies and relationships. Such complexities in computationally intensive science pose great challenges for both research data management and research reproducibility. On the one hand, there are data sources and codes as well as their 'footprints' to be captured and documented, which is often time-consuming and not considered part of doing science. On the other hand, highly automatic processes of data generation and analysis make it even more important for scientists to be able to track an analysis all the way back to the raw data and original configurations for reproducibility.

Although metadata standards have been created in major disciplinary fields, they are mostly designed and used for describing the end products of a research lifecycle either in the form of datasets or publications. As such, information important for provenance purposes generated during the research lifecycle is often either missed or skipped and needs to be recreated after the fact. Provenance metadata and other documentations created in this style are not only prone to errors and inaccuracies, but also expensive due to the time and expertise taken to track them down and enter into the system. In addition, data and the parameters and procedures used in processing data vary greatly from discipline to discipline, hence the metadata standards designed for one discipline may not be suitable in another. This is the case for Gravitational Wave (GW) research data management.

GW research is computationally intensive and the pipelines constructed for analysis runs may be split, derived, or merged to produce new workflows to be used for execution on computing grids. GW data go through a series of processing (segmenting, calibration, and registering) stages before they can be used for analysis. Along the way metadata describing these processes, codes, and resulting outputs must be captured and made discoverable and accessible so that the data and code used in an analysis can be tracked, verified, shared, reused, and/or reproduced. For example, detector characterization is a process involving the use of parameters and algorithms to cross check data quality and ensure the accuracy and validity of data to be fed into workflows. The data generated from detector characterization would need to be able to tell the source (interferometric detector), sample rate, latency, time covered, number of channels, and so on, all of which are unique to this research field. Given the nature of GW research lifecycle and disciplinary idiosyncrasies, current metadata standards developed for end-product description do not offer the level of granularity nor the kind of semantics needed for tracking, verifying, and reproducing GW analyses.

To address this need, several questions must be answered first: 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? We recognize the role of current metadata standards such as the Resource Metadata for the Virtual Observatory (IVOA, 2007) and Astronomy Visualization Metadata (AVM) (Hurt, Christensen, and Gauthier, 2008) in describing astronomy data. However, the validity and usefulness of existing standards has yet to be proven through formal evaluation. It would be risky to blindly adopt any

 doi:10.2218/ijdc.v11i1.399

metadata standard or develop a metadata model for a complex research domain without an in-depth understanding of the needs for metadata.

Motivated by the questions raised above, we conducted interviews with GW researchers and examined the research artefacts (configuration files, code files, workflows, and outputs) as well as their dependencies and other relationships. This paper reports the findings from interviewing GW researchers and approaches used to develop a metadata model with a focus on provenance metadata and research reproducibility. The following sections will first review publications on provenance and reproducibility of scientific research and then describe the characteristics of GW research lifecycle and the impact of these characteristics on the desired metadata functions. A summative report of the themes that emerged from the interviews will provide evidence to support the metadata function framework. The last section will discuss a framework of metadata for reproducibility of scientific research based on our project experience.

Relevant Literature

Sharing and reuse of research data is a practice adopted by many scientific communities today. In order for data to be sharable and reusable, they must be good quality, verifiable and discoverable. The quality and verifiability requirements for research data in many ways are associated with provenance metadata, which underpins research reproducibility.

The concept of reproducible research has been gaining attention from research communities, as data-driven science is becoming the norm over the last couple of decades. Reproducible research is considered as 'the practice of distributing, along with a research publication, all data, software source code, and tools required to reproduce the results discussed in the publication' (Shulte, Davison and Dye, 2013). This means that reproducible research requires the support of the methods for performing, preserving, and transmitting research, methods for storing analysis of data, and tools for storing papers and performing analysis (Thompson and Burnett, 2012).

Whether research can be reproduced is determined by many factors. Scientists' behaviour and practices in conducting science, the documentation of data, methods, and procedures, and the infrastructures supporting such documentation can all affect the reproducibility of research. For research to be reproducible the results must be verifiable, that is, datasets and code must contain information necessary for tracking, quality control, and reuse purposes. However, creating necessary documentation for data and code to be verifiable and reusable takes a great deal of effort and time, which has been cited as the 'biggest barrier' for scientific data sharing and reuse (Stodden, 2010).

Depending on the nature of the research, there may be different kinds of reproducibility. Empirical reproducibility, as Stodden (2013) puts it, refers to 'the traditional scientific notion of experimental researchers capturing descriptive information about (non-computational) aspects of their research protocols and methods.' Computational reproducibility emphasizes verifiable computing results that facilitate search, amalgamation, and tweaking of data and code (Gavish and Donoho, 2012).

Metadata describing data, code, and results is considered as critical in the reproducibility of research because it provides the provenance 'information that helps determine the derivation history of a data product, starting from its original sources'

(Simmhan et al., 2005). In scientific computing, provenance is considered as a process in which 'all the derivations, datasets, parameters, software and hardware components, computational processes, digital or non-digital artefacts' were used to derive and influence the data product (Deelman et al., 2010). Provenance metadata therefore documents the history of how a data product came into being. Many functions of data management, such as controlling data quality, tracing audit trails, replicating data and results, attributing to creators and contributors, and discovering data, are dependent on the provenance metadata.

Provenance metadata has been studied extensively and a good deal of literature has been published on the subject. Bose and Frew (2005) conducted a thorough review of standards, types of data processing, techniques, and technologies related to data lineage, otherwise known as data provenance. Workflow management tools such as Pegasus are being used and augmented to address the provenance questions (Miles et al., 2008). The Open Provenance Model developed by an international collaboration led by Luc Moreau (2010) defines three nodes – Artefact, Process, and Agent – and a set of dependency relationships and roles in its abstract model. While it is designed to be generic and applicable in computational intensive science fields, it is yet to be tested for its applicability in specific science research domains.

The reproducibility of research relies on metadata that documents the code, methods, parameters, data processing, and other artefacts important for provenance purposes. It is this type of metadata that makes the assessment of data quality and validity possible for data sharing and reuse.

The GW Data Flow

The Laser Interferometer Gravitational-Wave Observatory (LIGO) consists of two separate installations of interferometers in the U.S. and another one in Hanover, Germany. The Virgo interferometer is another installation located in Italy. These interferometric detectors are constructed with sophisticated engineering and technology and generate terabytes of data every day. Data generated from the detectors are transferred to a network of supercomputers for storage and archiving. Once the data are secured, scientists can use computer programs to process and analyse the data (LIGO Laboratory, 2016).

The fact that all the processing and analysis work requires computer programs to perform makes provenance information critical for data tracking, verifying, sharing, and reuse in this computationally intensive big science, big data research field. From a data flow point of view, the provenance metadata for GW datasets needs to consider both the pre-processing and the in-analysis processing. Provenance metadata generated during pre-processing stage is relatively straightforward, which Brown et al. (2006) has a detailed description. At the in-analysis processing stage, the GW analysis is performed through constructing and running pipelines that specify the various steps and sequence of execution of computation. For example, in searching for an inspiral signal, the output from one or two laser interferometric detectors is examined for signals of particular shape. This shape is called a template that specifies the process of analysis (Rana, 2006). Typical inspiral templates contain three major components: process, data, and parameters (Figure 1). Each of the components contains information that can be modelled as provenance metadata. Figure 2 shows three blocks of metadata for researchers to identify the process and specify the input data source. They address

 doi:10.2218/ijdc.v11i1.399

questions such as who created this process, which program it used, what the time boundaries of data source were, and where the output should be sent.

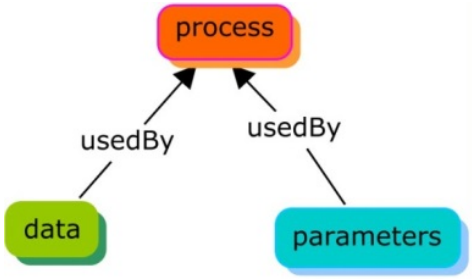

Figure 1. Components and their relationships in a GW TEMPBANK file.

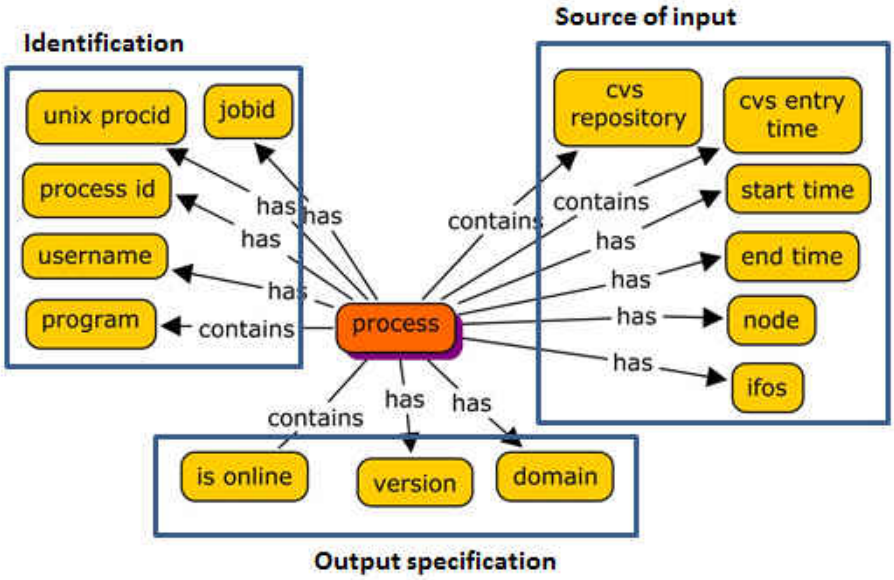

Figure 2. Provenance metadata for the process component in a template.

The ways that GW workflows are constructed make the provenance metadata inherently complex and challenging. As Brown et al. point out, 'workflows may be parallelized by splitting the full parameter space into smaller blocks or parallelizing over the time intervals being analyzed. The individual units are chained together to form a data analysis pipeline. The pipeline starts with raw data from the detectors, executes all stages of the analysis, and returns the results to the scientist' (Brown et al., 2006). Although analyses of GW workflows and pipelines will help to model the metadata for managing the data and reproducing results, it is necessary to understand scientists' practices in documenting and managing data, code, and output of analyses to generalize requirements and represent such requirements accurately in metadata models.

Gathering User Requirements

The LIGO Scientific Collaboration (LSC) community has established data practices and infrastructure for managing the data and code, as well as results from data processing and analysis. Understanding scientists' data practices and priorities is necessary for two reasons. First, producing verifiable computing results requires subtle adjustment to the work habits of scientists (Gavish and Donoho, 2012). Second, any adjustment to their work habits and practices must be based on a thorough understanding of research workflows, data flows, and priorities at each stage of the research lifecycle. The process of obtaining such an understanding is used to gather user requirements for the metadata model.

A number of sources are available for gathering initial information, including the LSC-Virgo website and wiki pages of working groups, configuration files and corresponding intermediate outputs and final results, metadata descriptions at the LSC document centre, and documentations for databases. We also had meetings with scientists to gather input for designing a formal interview protocol. These sources were carefully examined and analysed to derive concepts and components involved in GW research lifecycle, which are categorized based on their roles and uses in the GW research.

We also conducted a formal interview with eight members of the LSC community. The interview was designed to:

- Understand research data and analysis needs and habits of GW scientists;
- Learn about the requirements for discovering, tracking, documenting, and archiving data and analysis products, including workflows, input data, intermediary and final data products, software and its versions, as well as other computational artefacts;
- Define types of entities and relationships among the entities, as well as current systems used for identification management of these entities; and
- Collect vocabularies, including category lists, terminology for instruments, parameters, data status and operations, analysis methods and techniques.

The first group of questions collected information about the interviewee's status, experience, involvement in GW research, and motivations for keeping documentation. Each interviewee was asked to answer following questions:

1. What is your current position, and how long have you been in it?
2. How long have you been in your field of study?
3. Please briefly describe your role in the research group.
4. What motivates you to keep good documentation on the data and workflows you generated/used?

The second part of the interview asked the subjects to describe their work habits and practices through the following questions:

1. Please briefly describe the process of conducting an analysis.
2. How do you usually start an analysis project?

3. What are the most useful search options when you need to find data and/or workflows?
4. If you were to reproduce the same analysis, how much effort does it take to perform it?
5. What would you need to know in order to trust that you could reuse someone else's workflows?
6. Do you usually make extra documentation on your analysis data or products besides the system provided information about analysis input, output, parameters, etc.?
7. What aspects of managing data analysis workflows take most of your time?
8. If you are to rank the importance of a workflow management system, what would the top three features be?
9. In your mind, what should a data management system facilitate for your research a) when starting a project? b) during the project? c) at the end of the project?

The eight interviewees were recruited on site at the LSC-Virgo meeting during March 16-18, 2015. Recruitment took place through convenience and snowball sampling. Eligibility criteria included active LSC-Virgo membership, and particularly, affiliation with the Compact Binary Coalescence (CBC) subgroup. Since the priority of the interview was to identify emerging concepts and validate those that have been identified earlier from research on GW data and workflow samples and wiki pages, final subject selection included consideration of the interviewee's role and specialty in order to cast a wide coverage of different areas of GW research. IRB approval was sought and obtained before the start of the interviews.

Participants' backgrounds included academic degrees in physics, including areas like particle physics and gravitational physics. All participants were currently associated with academic institutions: four with institutions in the United States, and four with institutions in Germany. Two of the participants were professors at their institutions, while five were post-docs and one was a graduate student. All had years of experience with LSC, ranging from four years to 15 years, with an average of eight years. Within the LSC, participants held various and sometimes multiple roles. Three held leadership roles (S1, S2, S4), and were responsible for heading specific subgroups. Three were search specialists (S5, S6, S7), focusing on a particular type of search[1]. Two were responsible for developing code and pipelines (S3, S7), two were focused on data analysis (S2, S8), and two performed detector characterization work (S5, S6). All eight participants were male.

Each interview session ran for approximately 30 minutes and was audio recorded. The audio from the interviews was transcribed and, along with interviewer notes, was coded by two researchers with concepts derived from the aforementioned sources. The codes were modified and updated as new concepts and categories emerged from the coding process. The code list from two researchers were compared and merged to produce a list of themes. Each of the eight interviewees was given a code, ranging from S1 to S8.

Three major thematic areas emerged from the coding process: motivation for keeping documentation, reproducibility, and metadata that is against the grain. For each

1 Note that the term 'search' in this context refers to the search for gravitational wave in the data generated from interferometric detectors. It is synonymous to data analysis.

theme, several significant aspects are addressed below. Responses from each of the participants were considered in our exploration of each aspect, and one or two representative quotes are presented where possible.

Findings

Motivation for Keeping Good Documentation

Keeping good documentation of the data, code, parameters, pipelines and the changes made over the course of analysis is motivated by trackable, verifiable, and reusable results, which are all important links in the reproducibility chain. This sentiment resulted from several factors:

Trust

When a piece of code or a pipeline is to be reused, researchers need to first verify the code or pipeline to make sure it works in the way it is designed to run and turns out the results as documented.

> '…a page somewhere that documents what they did. And then for the code that they used, verification of that has been checked-off by a committee that looked it over. And it's not actually just sufficient for the committee to make the check mark, personally I would want to know that the tests that they did were sane…The best thing to do is have unit tests that check that this does what it does' (S2).

Accountability

This concept has two meanings: one is closely synonymous to reproducibility, as one interviewee referred to as 'exactly how you did what you did' and the other is related to accountability to the funding agency. The first meaning is most relevant to the internal community, while the second meaning holds relevance for external parties.

> 'It's very much driven by the fact that we're funded by the government, and the NSF now demands it, as well they should, and demands that the data be released with full documentation, and demands that every paper that we write, in addition to just the paper itself, words on paper, you should also archive the data that went into the results that are presented' (S4).

Continuity

As with any research and educational institution, there is a constant change in GW research group members because new graduate students and post-doctoral researchers regularly enter into and leave a group due to educational and funding cycles. Documentation supports continuity among a dynamic community membership.

> 'It's always the new people, I guess, that you're interested in documenting it for. We regularly have a turn-around of both students and post-docs coming in and out and if they want to get involved in running pipelines, in looking at data and understanding things' (S8).

 doi:10.2218/ijdc.v11i1.399

Reproducibility

As a field of research that employs intensive computation and big data, almost every step in GW research involves some sort of programming code. As soon as the data are generated from the interferometric detectors, a series of processing is performed to calibrate, segment, and frame data, all of which are controlled by computer programs. A GW search analysis also deploys a large amount of code for parameter selection, data input and output specification, error handling, pipeline generation, and computing job scheduling and monitoring. The reproducibility of GW research is therefore essentially the reproducibility of the code. In this sense, it is exactly what Gavish and Donoho (2012) describe as computational reproducibility. From our interviews, we identified several themes that are useful for building the metadata model for reproducibility.

Code dependencies

The code used for a run may be stored in a GitHub repository and/or a file directory, which is then linked to a wiki page that contains all information and output from the run. Because a run often uses a large number of program codes and they are modified and changed frequently, it is important for anyone who is to reuse the code to know the dependencies between the codes in order to reuse them properly.

> 'And if everything works, what you do is you go in, login to the cluster. You go check out that Git hash. You build the code. That's the first terrifying step, because the code usually won't build, just because the code has so many dependencies… we write down the Git hash of the code we use, but this code depends on a lot of other codes' (S7).

Code dependency is especially important when an 'old' analysis needs to be reproduced:

> 'If it's that old, that will tell us the exact version of the code that was installed. We can check this version out and install it. If the code version is particularly old, we may have to install older dependencies as well' (S8).

Code versions

Interviewees described how important it is to know the code version in order to reproduce an analysis run. Code version information is often stored in a GitHub repository created and maintained by a group or researcher, and is important evidence for determining components necessary to reproduce a run.

> 'It depends on the amount of diligence you put into in the first place. For some – I know that – so when I first did searches as a grad student, I probably would not be able to reproduce that analysis if I tried, for many reasons. The code has changed, [and] the data has been shifted around and/or does no longer exist in easily accessible places. As I mentioned before, some of the data quality information has been either lost or no longer present somewhere' (S5).

Documentation

The importance of documentation was commonly recognized and the common practice of documenting is through creating wiki pages. The following quote tells what it is important to document and the challenges in retrieving specific information from the current wiki-based system of documentation:

> '…if we want to go back to an analysis that was done in 2006 and get the raw data, that went into it – not the raw detector data, which is all archived, we can get that. But the intermediate analysis data, like things that we call triggers where we shift through our time series of data and then look for excursions … He probably put some information about it in the wiki back in 2006. Those wikis are still there, we can dig them out but it can be very difficult and laborious to do that...' (S4).

Content reproducibility vs. code reproducibility

Reproducing the science content does not have to use the exactly same programming language. The same results may be reproduced by using another programming language different from the original code used to produce the results. Thus, GW researchers distinguish between different types of reproducibility:

> 'And it's not as easy to reproduce something exactly when you've got so many moving pieces. So it's much easier if someone produced an analysis simply by compiling some C code and writing everything from scratch and then putting a static binary somewhere, then yes, you could probably reproduce everything almost exactly but that's almost never what we do anymore' (S5).

Verification

Reproducing an analysis is not as simple as taking someone's code and rerunning it in the hopes of producing the same results as it did before. Codes and data must be verified to make sure the analysis is 'sane.' Wiki pages play a key role in obtaining the information about code, such as versions and dependencies.

> 'Trust someone else's data: need to verify code and data. There could be a number of possibilities: analysis by a different group, using a different algorithm that affects the verification and reproducibility' (S1).

> 'We usually start checking by looking at the wiki page. Need to make sure the wiki page is not broken' (S1).

Metadata that is Against the Grain

During the lifecycle of GW research, metadata seems to be critical at a number of points: 1) when raw data have been calibrated and segmented and entered into a state of readiness to be used in a search, 2) when a workflow has been constructed and pipeline

 doi:10.2218/ijdc.v11i1.399

generated, and 3) when a run is completed. At each of these stages there are different properties and relationships to be represented in a metadata model, many of which are not available in any of the existing metadata standards.

Data calibration and preparation

Prior to the start of a search workflow, raw data has been collected from the interferometers, as well as detector characterization data which is used to assist in interpretation. Metadata on the data sample rates and channels, the interferometers, and the data quality flags are all recorded. At the time of a search, this metadata is needed to help identify, request, and relate correct datasets from both sources. Accuracy is particularly important given the large amount of data being generated.

> '…that's the raw data and then there is all this, like I call, intermediate data, like lots of triggers and things. Quite voluminous, but it's not petabytes. It's tens, hundreds of gigabytes and that's sitting on discs somewhere in our clusters, and you can find it if you know where to look' (S4).

Pipeline generated

A specific pipeline or workflow bears relationships to many other entities in LIGO work, including raw data, characterization data, and executables. Pipelines are created by a researcher for a specific type of search, and many properties of a particular pipeline are recorded in a configuration file written by the researcher. As entities, pipelines are highly interrelated within the metadata model, and thus depend on properties from a number of other entities. Given the complicated nature of the workflows, metadata is needed to help researchers fully identify the vast contents of a pipeline.

> '…they have to write a pipeline that goes through the data and looks for something or analyses something. That pipeline makes you some software, some of which has already been written, some of which you have to write. If you look at the LIGO scientific collaboration software repository, you'll see tens of millions of lines of code' (S4).

Run completion

When a run has been completed, output data is generated and must be identified and related to the originating workflow. At the same time, a vast number of intermediate data files have also been produced. These must be accurately located and related to particular processes within a workflow using accurate metadata to assist in result interpretation if needed.

> 'At the end, things should be saved so that everything that was done is accessible. That any choices that were made at the time of which data to analyze get recorded' (S6).

Discussion and Conclusion

As related by the interviewees, tracking components of an analysis project in the current environment poses many challenges. Pipelines contain a complex series of interdependent executables, each producing intermediate files that may be valuable in verifying or reproducing a run, but which may not be easily accessible. In addition, understanding the executables utilized in the various steps of an analysis may require researcher to investigate through documentation and online code storage. The inclusion and utilization of provenance metadata at various stages in the research cycle can improve workflow tracking as well as reproducibility. Readily available provenance metadata concerning executable code, such as its creator, version, computing environment, and location could facilitate researchers in understanding and reproducing the steps in a workflow.

Metadata concerning output and intermediate files is also crucial in understanding an analysis project. Researchers may need to know the location of the files, as well as any workflows, specific processes, and timestamps associated with them. Throughout the interviews, participants stressed the importance of the configuration file, a file serving as base instructions for the compilation of a workflow. To reproduce a workflow, researchers need the original configuration file associated with it. The configuration file relates the workflow to data sources and executables. Describing the configuration file with appropriate metadata, such as its creator and purpose, and capturing its relationship to workflows and other entities provides a crucial link in tracking analysis output back to its original data sources and parameters.

Interview results also reveal the metadata types and functions that GW researchers consider important in supporting their work. Metadata concerning the researchers themselves and their relationships to workflow inputs and data products seems crucial in determining trust and reliability. As such, this metadata is important in enabling verification and reproducibility. Determining the originator of a configuration file, the creator of an executable's code, or the owner of an intermediate file directory represent important researcher tasks that metadata should be able to support. Tracing the results of an analysis back to a configuration file or a piece of code represents another type of important task that participants described. Metadata should enable a specific workflow instance to be uniquely identified and associated with any of its data products. At the same time, metadata about the workflow should capture its relationships to configuration files and executable codes, thus allowing a provenance path from workflow products back to workflow components. As a vast amount of code may be used in any specific workflow, tracking and understanding code and how it may have changed over time is another important group of tasks described in the interviews. Metadata including location, version, and creator may be helpful in accessing, altering, and utilizing any code of interest.

Overall, findings from the interview process support the belief that GW research, like other computationally-intensive scientific research, involves complex, interdependent workflows. Provenance data needs are high during the entirety of the research lifecycle, and particularly so at certain stages. In considering reproducibility, GW researchers distinguished between reproducing code and reproducing content; having access to the appropriate metadata facilitates in either scenario allows researchers to make comparisons between computational environments even if they are unable to fully recreate them. Identifying all inputs, outputs, and processes, their provenances, and the relationships between them is challenging but vital for this

reproducibility. The metadata needs within this community display a level of granularity and specificity not provided by existing metadata models. As such, a metadata model specifically for GW research within the LIGO community is justified.

This metadata model is currently under development based on the user requirements generalized from our review of LSC research artefacts and interviews with LSC members. From a metadata perspective, functions of tracking and discovery rely on the availability of consistent identification and linking between dependency relationships throughout the system. The challenges in this metadata modelling lie in maintaining a balance between the completeness of metadata representation and the cost of scientists' time in doing so. An iterative modelling process with feedback from GW scientists will help maintain the balance.

Acknowledgements

This project is supported by the U.S. National Science Foundation grant #ACI-1443047.

References

Bose, R. & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Survey, 37*(1). doi:10.1145/1057977.1057978

Brown, D.A, Brady, P.R., Dietz, A., Cao, J., Johnson, B., & McNabb, J. (2006). A case study on the use of workflow technologies for scientific analysis: Gravitational wave data analysis. In I.J. Taylor, E. Deelman, D. Gannon, and M.S. Shields (Eds.), *Workflows for e-Science*. Berlin: Springer-Verlag.

Deelman, E., Berriman, B., Chevenak, A., Corcho, O., Groth, P., Moreau, L. (2010). Chapter 12: Metadata and provenance management. In A. Shoshani and D. Rotem (Eds.), *Scientific Data Management: Challenges, Existing Technology, and Deployment*. CRC Press. Retrieved from http://arxiv.org/ftp/arxiv/papers/1005/1005.2643.pdf

Gavish, M. & Donoho, D. (2012). Three dream applications of verifiable computational results. *Computing in Science and Engineering 14*(4). doi:10.1109/MCSE.2012.65

Hurt, R., Christensen, L.L., & Gauthier, A. (2008). Astronomy visualization metadata (AVM) standard – version 1.2 rc1. Retrieved from http://www.virtualastronomy.org/AVM_DRAFTVersion12_rlh02.pdf

IVOA. (2007). Resource metadata for the virtual observatory version 1.12. IVOA Recommendation 2007 March 2. Retrieved from http://www.ivoa.net/documents/REC/ResMetadata/RM-20070302.html

LIGO Laboratory. (2016). LIGO technology. Retrieved from https://ligo.caltech.edu/page/ligo-technology

Miles, S., Groth, P., Deelman, E., Vahi, K., Mehta, G., & Moreau, L. (2008). Provenance: The bridge between experiments and data. *Computing in Science and Engineering, 10*(3). Retrieved from http://eprints.ecs.soton.ac.uk/20874/1/cise2008.pdf

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., & Van den Bussche, J. (2010). The open provenance model core specification (v1.1). *Future Generation Computer Systems.* Retrieved from http://eprints.ecs.soton.ac.uk/21449/1/opm.pdf

Rana, O.F. (2006). Gravitational wave analysis. Retrieved from http://www.gridprovenance.org/publications/CardiffUseCase2.pdf

Schulte, E., Davison, D., & Dye, T. (2013). Reproducible research [Documentation]. Retrieved from http://orgmode.org/worg/org-contrib/babel/intro.html#reproducable-research

Simmhan, Y.L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-Science. SIGMOD Record, 34(3). Retrieved from http://www.cs.indiana.edu/dde/papers/simmhanSIGMODrecord05.pdf

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences. MIT Sloan research paper no. 4773-10. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193#%23

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. *IMS Bulletin, November 17*. Retrieved from http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/

Thompson, P. A. & Burnett, A. (2012). Reproducible research. *CORE Issues in Professional and Research Ethics*, *1*(Paper 6). Retrieved from http://nationalethicscenter.org/content/article/175