Modeling Domain Metadata beyond Metadata Standards

Jian Qin[1], Brian Dobreski[1]
[1]Syracuse University

Abstract
The Laser Interferometer Gravitational-wave Observatory (LIGO) project to detect gravitational waves represents a complex, distributed scientific endeavor posing specific challenges for reproducibility and data management. The integration of provenance and other metadata information into the workflow stands as one means of addressing such challenges. The goal of a metadata model for the LIGO workflow is the provision of metadata describing all the data products at each significant milestone in the data analysis pipeline. Given both the highly specific domain and the need to support current analysis tools, the development of such a model demands a more complex, comprehensive approach. For this reason, we pursued a multipronged approach to metadata modeling, gathering users' conceptions, system information, research artifacts, and other organizational documents, and worked to combine the findings into one final model. This approach provided a thorough understanding of the overall research lifecycle and insight into scientific workflow metadata modeling.
Keywords: Domain metadata; Metadata modeling; Provenance metadata; Gravitational wave data management
doi: 10.9776/16563

Acknowledgements: This project is supported by NSF award #ACI-1443047.
Contact: jqin@syr.edu, bjdobres@syr.edu

1 Introduction

The Laser Interferometer Gravitational-wave Observatory (LIGO) is an ambitious project to detect gravitational waves from astrophysical sources such as colliding neutron stars, black holes, and supernovae (Deelman et al., 2006). Gravitational wave research itself is computationally intensive, involving a number of data sources and products that may be processed, split, merged, or derived through applying different workflow programs. This requires complicated analysis pipelines, containing hundreds of thousands of discrete computational tasks, with complex dependencies and correspondingly large numbers of intermediate data products. Such pipelines are time consuming to run and pose challenges to reproduction (Stodden, 2010). For example, sharing code and intermediate data files among researchers is necessary to allow data and code reuse, but currently requires additional manual coordination, and is error-prone. At the same time, data reuse and pipeline reproduction is growing more important for the LIGO community given the increasingly complex and distributed nature of the workflow, as well as concerns about verification of potential gravitational wave events. The integration of provenance and other metadata information into the workflow stands as one means of facilitating the reproducibility and management of LIGO data and analyses.

The development of a metadata model for gravitational wave research represents the first part in a multicomponent research process concerning LIGO workflows, with the ultimate goal of advancing the systems and processes in use by gravitational wave researchers. In this initial stage, research will create a comprehensive metadata model for the pipelines in use by the Compact Binary Coalescence (CBC) Group, a specific analysis subgroup within LIGO focusing on coalescing compact binaries, one of the most promising sources of gravitational radiation (Abadie, 2010; Abadie, 2011). The goal of this model is the provision of accurate and complete metadata describing all the data products at each significant milestone in the CBC data analysis pipeline. This workflow-aware metadata model will capture not only essential attributes of data and computational artifacts but also the important relationships between them. The resulting metadata should be expected to better support tracking, curation, access, and validation of data sets, with the ultimate intent of integrating into and supporting developing tools that will facilitate the LIGO analysis workflow. Previous work examining the development of scientific metadata has noted a number of particular challenges, including heterogeneous data sets, derived data products, work and data distributed across networked environments, and the need for robust provenance metadata (Jones et al., 2001; Deelman, 2008). Given these challenges,the highly specific domain, and the need to support pre-existing tools in the analysis process, the development of a LIGO pipeline metadata model demands a more complex, comprehensive approach, focused specifically on the intended users and their environments. For this reason, we pursued a multipronged approach to metadata modeling, gathering

users' conceptions, primary sources of workflow documentation, and other organizational documents, and worked to combine the findings into one final model.

2 Gathering User Requirements and Research Artifacts

Prior to any data gathering, we first reviewed general user requirements implicit in the research premise. In order to achieve effective and efficient reuse and reproducibility, LIGO scientists must be able to discover and reuse previous analyses and their data products. As such, retrieval functionality, integrating traditional object-based retrieval based on metadata, is an important consideration. Users need to be able to easily catalog their computational results, and be able to search for and reuse previously derived data and previously constructed analysis with high precision and recall. Given the complex and automated nature of their workflows, users also need automatic collection of metadata and parameter information generated by workflows running in a distributed environment. A key component to supporting this is the assignment of unique and persistent identifiers to data objects when they are produced and the ability to reliably establish the provenance of each object through metadata. Such digital object identifiers are widely used across disciplines (Tilmes & Yesha, 2010). Finally, any metadata structures recommended must be compatible with existing LIGO workflow management software. The CBC Group currently employs Pegasus (Cadonati, 2009; Aylott, 2009), a workflow management system that allows scientists to specify their workflows at a high-level of abstraction, and carries out this plan in a distributed environment. With these basic requirements in mind, we first turned to the LIGO scientists themselves to gain a better understanding of both the needs and conceptions of these users.

We conducted a series in-person interviews during the March 2015 LSC-Virgo Meeting in Pasadena, California. Attendees of the conference represent LSC and Virgo participants from institutions around the world. Eight participants were recruited to take part in the study due to their involvement with CBC Analysis Group, who ranged from graduate students to professional astrophysicists with over 20 years of experience. All participants were male, and all were associated with institutions in the United States or Germany. During the interview, participants were asked to respond to two main groups of questions. The first set of questions concerned participants' background, including their general experience, their current involvement in LIGO, and their documentation motivation and habits. The second set of questions addressed specific work habits and practices. Participants were asked to share their views on the analysis process, information retrieval needs, analysis reuse and reproduction, and workflow management system needs. Interviews lasted approximately 30 minutes and were audio recorded. Following the interviews, interviewer notes and transcriptions of the audio recordings were open coded by two coders for major themes including the general concepts, processes, relationships, and vocabularies in the CBC workflow, as well as beliefs and practices concerning data management and documentation.

In addition to initial user requirements and interview data, we also reviewed various forms of LIGO documentation for further insight into the workflow and its products. This included the website for the LIGO Scientific Collaboration, wiki pages of the LIGO working groups, metadata descriptions from the LIGO Scientific Collaboration Document Center, and documentation for the Pegasus workflow management system. Samples of actual research artifacts were consulted as well, including configuration files, intermediate outputs, and final results. These sources were carefully examined and analyzed to derive concepts and components involved in the gravitational wave research lifecycle.

Results from this multipronged data collection were reviewed and compared at a broad level. In general, interview results were helpful in determining high level objects and processes in the CBC LIGO workflow, while user requirements and documentation analysis revealed relationship and identification needs. The presence of common entities in all data sources allowed clarification of their relative importance. For example, the configuration file emerged as a central component of the overall workflow. The most commonly referenced topics included documentation sources (LIGO wiki), software (Pegasus Workflow Management system), input sources (configuration files), and output (analysis result web pages). Through an iterative process of mapping, the most important broad topics andrelationships were arranged into a high level conceptual map (Figure 1), which was used to identify the specific portions of the research life cycle of interest to a metadata model. Entities of interest and other sources of metadata were identified and set aside for potential inclusion in the metadata model.

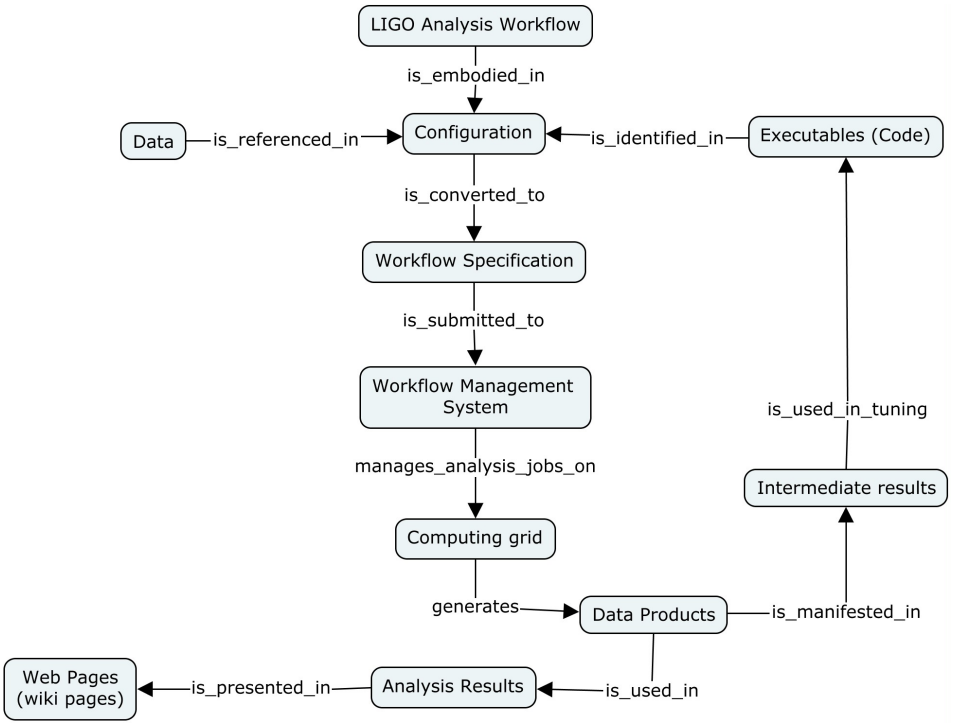

Figure 1. A high level view of LIGO analysis workflow

3 Conclusion

Developing a metadata model for a highly specific, scientific domain required a distinct, multipronged approach. Information from users, systems, and research artifacts was needed to provide a full understanding of the CBC Group's workflows and metadata needs, as well as to triangulate findings. Through examining multiple sources, the relative importance of entities became clearer, along with their pertinent relationships and attributes. Most importantly, the varied perspectives considered provided a thorough understanding of the overall research lifecycle, allowing us to locate the most vital aspects of the workflow needed for reproducibility and data management.

The metadata model generated from the approach is still in the process of being improved as the Advanced LIGO is entering full operation. With the anticipation of larger amounts of data, code, and other research artifacts generated by the much more powerful instrumentation in Advanced LIGO, this metadata model will play a significant role in supporting the broader goal of advancing LIGO's work. The metadata dictated by this model will better enable scientists to discover data and create data analyses, while improving the accessibility, usability, reproducibility, and reusability of LIGO work. At the same time, the modeling approach presented here provides a unique contribution to the methods in identification of science requirements, scientific metadata modeling, and eScience workflow management.

Acknowledgement

This research is supported by NSF grant #ACI-1443047.

References

Abadie, J., Abbott, B. P., Abbott, R., Accadia, T., Acernese, F., Adhikari, R., ... & Bizouard, M. A. (2010). Search for gravitational-wave inspiral signals associated with short Gamma-Ray Bursts during LIGO's fifth and Virgo's first science run. The Astrophysical Journal, 715(2), 1453.

Abadie, J., Abbott, B. P., Abbott, R., Abernathy, M., Accadia, T., Acernese, F., ... & Beveridge, N. (2011). Search for gravitational waves from binary black hole inspiral, merger, and ringdown. Physical Review D, 83(12), 122005.

Aylott, B., Baker, J. G., Boggs, W. D., Boyle, M., Brady, P. R., Brown, D. A., ... & Pretorius, F. (2009). Testing gravitational-wave searches with numerical relativity waveforms: results from the first Numerical INJection Analysis (NINJA) project. Classical and quantum gravity, 26(16), 165008.

Cadonati, L., Aylott, B., Baker, J. G., Boggs, W. D., Boyle, M., Brady, P. R., ... & Pollney, D. (2009). Status of NINJA: the numerical injection analysis project. Classical and quantum gravity, 26(11), 114008.

Deelman, E., Livny, M., Mehta, G., Pavlo, A., Singh, G., Su, M.-H., … Wenger, R. K. (2006). Pegasus and DAGMan From Concept to Execution: Mapping Scientific Workflows onto Today's Cyberinfrastructure. In High Performance Computing Workshop (pp. 56–74).

Deelman, E., & Chervenak, A. (2008, May). Data management challenges of data-intensive scientific workflows. In *Cluster Computing and the Grid, 2008. CCGRID'08. 8th IEEE International Symposium on* (pp. 687-692). IEEE.

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *Internet Computing, IEEE*, *5*(5), 59-68.

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

Tilmes, C. & Yesha, Y. (2010). Provenance artifact identification in the Atmospheric Composition Processing System (ACPS). https://www.usenix.org/legacy/event/tapp10/tech/full_papers/tilmes.pdf