7th International Digital Curation Conference
December 2011

Golden-Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository

Practice Paper

Paolo Missier, Newcastle University, UK
Bertram Ludäscher, Saumen Dey, Michael Wang, Tim McPhillips, UC Davis, USA
Shawn Bowers and Michael Agun, Gonzaga University, USA
Ilkay Altintas, UC San Diego, USA

July 2011

Extended Abstract

Experimental science is not a linear process. As we have noted in our recent prior work (2010a), publishable results routinely emerge at the end of an extended exploratory process, which unfolds over time and may involve multiple collaborators, who often interact only through data sharing facilities. This is particularly apparent in e-science, where experiments are embodied by computational processes which can be executed repeatedly and in many variations, over a large number of input configurations. These processes typically encompass a combination of well-defined processes encoded as scientific workflows, e.g., in Kepler (2006a), Taverna (2007a), etc., or as custom-made scripts, operations that move data across repositories, etc.

Current implementations of e-science infrastructure are designed to support primarily the discovery and creation of valuable data outcomes, while result dissemination has largely been confined to "materials and methods" sections in traditional paper publications. Spurred in part by pressure from funding bodies, which are interested in maximizing their return on investment, the focus of e-science research is now shifting on the later phases of the scientific data lifecycle, namely the sharing and dissemination of scientific results, with the key requirements that the experiment be repeatable, and the results be *verifiable* and *reusable* (2009a). The notion of *Research Objects* (RO) has emerged in response to these needs (2010c). These are bundles of logically related artifacts that collectively encompass the history of a scientific outcome and can be used to support its validation and reproduction. They may include the description of the processes used (i.e., workflows), along with the *provenance traces* obtained by observing workflow execution.

Importantly, the view of the experimental process they provide is focused on a selected few datasets that are destined for publication, rather than on the entire "raw" exploration. As a result, such a view is a "virtual" one, in the sense that it represents a linear and uniform account of the research, obtained by sifting through a possibly large space of partial and possibly invalid intermediate results, which were generated at different times (possibly by multiple collaborators with different environments).

The project described in this paper is set in the context of the Data Observation Network for Earth (DataONE) NSF project[1]. The goal of the project is to support this *experiment virtualization* step of the scientific data lifecycle, which we view as a prerequisite to the creation of shareable Research Objects. We have termed the project *Golden-Trail*, to emphasize that our architecture enables scientists to generate a "clean" account of their most valuable findings (the "golden data"), out of many possible, often only exploratory, analysis paths.

A number of challenges are associated to this idea: Firstly, it requires that provenance traces generated from heterogeneous and independeny processes, possibly operating on different e-science infrastructures, be combined into a single trace that represents the virtual experiment. Two traces that represent the executions of processes A and B logically "join" on any dataset produced by A and consumed by B. Automating this join step, however, requires that the repositories used to hold the intermediate data be "provenance-enabled" (2010b), and that the space of data identifiers used by A and B be mapped to one another. In general, this requires an explicit curation step with the scientist's direct involvement.

[1] http://www.dataone.org

Secondly, scientists may find it difficult to explore provenance traces unless they can relate them to the original experiment design from which they were derived. Thus, a specification of the experimental process, i.e., the set of workflows involved, must accompany the traces if scientists are to carry out effective curation. With these considerations in mind, Golden Trail involves the following elements:

- A *provenance model* for describing the lineage of process-generated data. Unlike existing provenance models, like the Open Provenance Model (OPM) (2011a) which ignores the nature of the process that generated the data, our model includes the description of the process specification in addition to the data dependencies observed during process execution. Importantly, we aim to accommodate the most common workflow models that are in broad use in e-science, including Kepler, Taverna, VisTrails (2006), Pegasus (2008a), Galaxy (2010d), and eScience Central (2011b). We denote our model D-OPM, to indicate that it is a backward-compatible extension of the OPM;
- A *provenance repository* for storing the "raw" provenance traces obtained from multiple executions of one or more processes, which represent the actual exploratory phase of scientific investigation;
- A *user environment* for the semi-automated construction of virtualized accounts of an experiment. The environment consists of two components: (i) a query interface into the repository, by which the scientist can explore and visualize the space of available traces, guided by the process specification part of D-OPM, and (ii) a *curation interface* by which scientists provide the necessary mappings across data generated by different traces (an explicit data curation step).

Implementation and Experimental Testbed

We have implemented a prototype for the Golden-Trail provenance repository that is designed to be integrated with the main DataONE architecture (**Figure *1***).

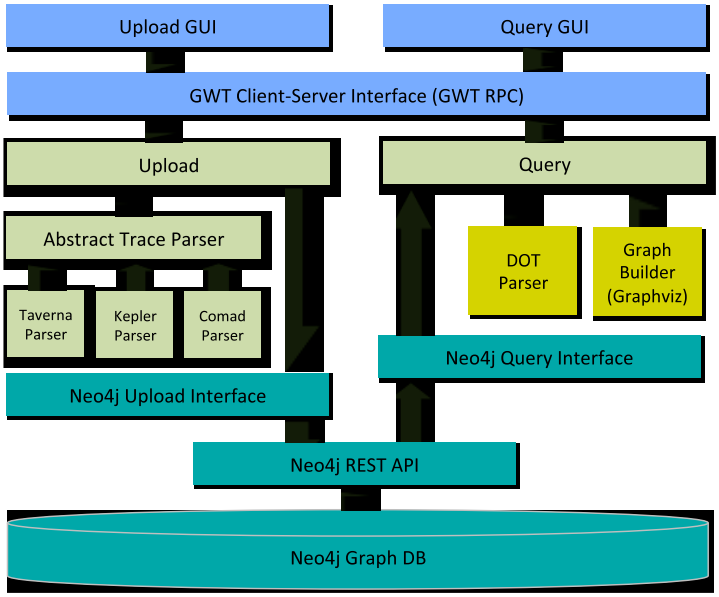

Figure 1. *Golden-Trail prototype architecture. The Neo4J graph database was used as a data layer as it naturally matches the graph data model used to represent provenance traces.*

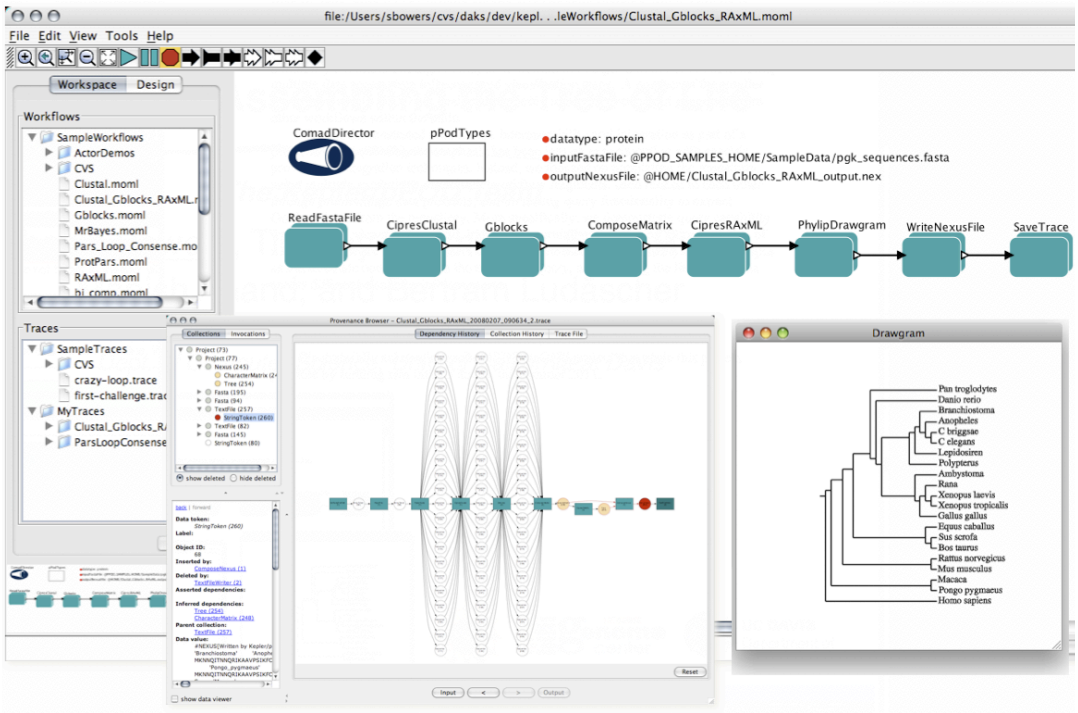

Figure 2: *Phylogenetics workflow (top) with provenance trace (bottom) from the Kepler/pPOD package.*

Our experimental testbed consists of a suite of pre-existing Kepler workflows, prepared from the "Tree of Life"/pPOD project (2008b). The pPOD testbed includes a suite of workflows for performing various phylogenetic analyses, using a library of reusable components for aligning biological sequences and inferring phylogenetic trees based on molecular and morphological data. The workflows are divided into various subtasks that can be run independently as smaller, exploratory workflows for testing different parameters and algorithms, or combined into larger workflows for automating multiple data access, tree inference, and visualization steps. A number of the smaller workflows within pPOD are designed explicitly to be run over output generated from other workflows within the suite.

Having demonstrated provenance interoperability and integration as part of a previous effort (2010b), the emphasis has been less on experimenting with specific provenance integration techniques. Instead, we focused on populating the repository using multiple executions of multiple workflow fragments, each related to each other through intermediate data products, and on testing query functionality to extract Golden-Trails from the repository. More specifically, we demonstrate query capability with different views of the result, including returning and rendering all or a portion of a run graph, where nodes represent whole workflow runs, and possibly with data nodes as intermediate connections, as the result of a query, emphasizing the lineage of data across different e-science infrastructures.

Acknowledgements

We gratefully acknowledge the NSF DataONE project that made this project possible by funding two interns during the summer 2011.

References

[journal article] (2011a) Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., et al. (2011). The Open Provenance Model --- Core Specification (v1.1). *Future Generation Computer Systems*, *7*(21), 743-756. Elsevier. doi:http://dx.doi.org/10.1016/j.future.2010.07.005

[report] (2011b) Hiden, H., Watson, P., Woodman, S., & Leahy, D. (2011). *e-Science Central: Cloud-based e-Science and its application to chemical property modelling*.

[proceedings] (2010b) Missier, P., Ludaescher, B., Bowers, S., Anand, M. K., Altintas, I., Dey, S., Sarkar, A., et al. (2010). Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science. *Proc.s 5th Workshop on Workflows in Support of Large-Scale Science (WORKS)*.

[proceedings] (2010a) Altintas, I., Anand, M. K., Ludaescher, B, Bowers, S., Crawl, D., Belloum, A., Missier, P., Goble, C., & Sloot, P. (2010). Understanding Collaborative Studies Through Interoperable Workflow Provenance. *Procs. IPAW 2010*. Troy, NY.

[proceedings] (2010c) Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., Roure, D. D., et al. (2010). Why Linked Data is Not Enough for Scientists. *e-Science (e-Science), 2010 IEEE Sixth International Conference on* (pp. 300-307). doi:10.1109/eScience.2010.21

[journal article] (2010d) Nekrutenko, A. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. Retrieved from http://dx.doi.org/10.1186/gb-2010-11-8-r86

[journal article] (2009a) *Nature, Special Issue on Data Sharing*. (2009). *Nature* (Vol. 461).

[journal article] (2008a) Kim, J., Deelman, E., Gil, Y., Mehta, G., & Ratnakar, V. (2008). Provenance trails in the Wings-Pegasus system. *Concurrency and Computation: Practice and Experience*, *20*, 587-597. doi:http://dx.doi.org/10.1002/cpe.1228

[proceedings] (2008b) Bowers, S., McPhillips, T. M., Riddle, S., Anand, M. K., & B.Ludäscher. (2008). Kepler/pPOD: Scientific Workflow and Provenance Support for Assembling the Tree of Life. *IPAW* (pp. 70-77).

[proceedings] (2007a) Turi, D., Missier, P., Roure, D. D., Goble, C., & Oinn, T. (2007). Taverna Workflows: Syntax and Semantics. *Proceedings of the 3rd e-Science conference*. Bangalore, India. doi:http://dx.doi.org/10.1109/E-SCIENCE.2007.71

[proceedings] (2006) Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. (2006). VisTrails: visualization meets data management. *Procs. SIGMOD* (pp. 745-747). doi:http://doi.acm.org/10.1145/1142473.1142574

[journal article] (2006a) Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., et al. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039-1065. John Wiley & Sons, Ltd. doi:10.1002/cpe.994