Presentation

Identifier Services: Tracking Objects and Metadata Across Time and Distributed Storage Systems

Maria Esteva
Texas Advanced Computing Center, USA
maria@tacc.utexas.edu

Ramona Walls
CyVERSE, USA
rwalls@cyverse.org

Abstract

Global identifiers are key to current and future access and reuse of data. Considering increasing data production, the complex and often messy nature of research data practices from which datasets are derived, and the ever-changing landscape of storage and publishing platforms, a single identifier type and a unique data location for a dataset does not function well nor scale. Instead, there is a need to use multiple identifiers throughout the lifecycle of a project, starting at the moment of data creation and well beyond publication to identify reuse. For complex datasets identifiers must accurately represent the diverse processes that generate the data. Thus, they must carry provenance metadata that describes these processes and make connections among their inputs and outputs. Despite the location, duplication, similarity, and archiving status of the data, its metadata must have a unified representation. These requirements have implications for implementation, including accounting for the validity of data over time, the technical resources that will support such infrastructure, and users' adoption.

Using real biology datasets, we are conducting investigations around Identifier Services (IDS). IDS is designed to bind dispersed data objects and verify aspects of their identity and integrity, independent of where the data are located and whether they are duplicate, partial, private, published, active, or static. IDS will allow individuals and repositories to manage, track, and preserve different types of identifiers and significantly improve provenance metadata of distributed collections at any point of their lifecycle.

One year into the research we have: (a) developed a generalizable data model (See figure 1) that maps genomic materials (e.g. specimen), processes (e.g. sequencing, alignment, experiments, analysis) and derived data to: global and or local identifiers and corresponding domain science (MIGS, INSDC) and citation metadata (DataCite); (b) used an API to automatically validate data associated to a global identifier and track their integrity, presence at an established location, and identity (similarity to an identical or similar dataset); and (c) implemented a user portal where the actions of the IDS are executed and its results recorded. The entities in the data model group files and metadata to corresponding processes, thus expressing their provenance. The portal provides landing pages for evolving representation of registered research projects where identifiers point to and from different data storage locations. We are using the data management infrastructure Agave (The Agave, 2016), which allows IDS to connect to repositories and access data to perform actions in a distributed computational environment. Bio-collection creators have been recruited to provide data and requirements for the prototype services, as well as structured feedback. Currently we can demonstrate a workflow in which users register their collection with IDS, select files and processes to reflect the provenance of a complex genomic dataset located both at a university storage resource and a centralized institutional repository, and conduct services across these different resources. We will report on the fitness of the data model to other science domains, provenance representation, access to the data, and the need for big data and metadata interface solutions.

Keywords: data modelling; provenance; data identifiers; distributed.

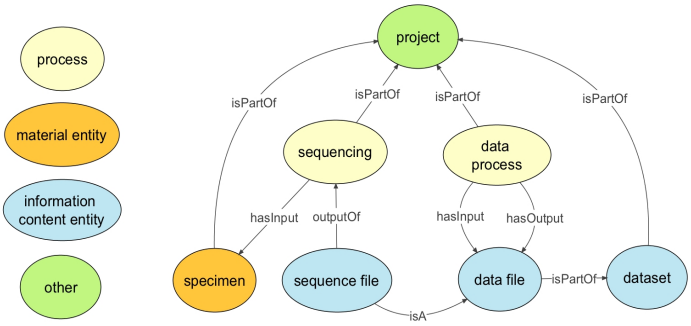

FIG. 1. Identifier Services (IDS) data model for Biology datasets adapted to Genomics.

Acknowledgements

This work is supported by the NSF EAGER: Collaborative Research: Evaluating Identifier Services for the Lifecycle of Biological Data. #26100741

References

The Agave Platform. (n.d.). Retrieved July 13, 2016, from http://agaveapi.co.

DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved January 10, 2007, from http://www.dublincore.org/documents/1998/09/dces/.

Heery, Rachel. (2004). Metadata futures: Steps toward semantic interoperability. In Diane I. Hillmann & Elaine L. Westbrooks (Eds.), Metadata in practice (pp. 257-271). Chicago: American Library Association.

Hillmann, Diane. I., Stuart A. Sutton, Jon Phipps, and Ryan J. Laundry. (2006). A metadata registry from vocabularies up: The NSDL registry project. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006, 65-75.

Lagoze, Carl, Dean Krafft, Sandy Payette, and Susan Jesuroga. (2005, November). What is a digital library anyway, anymore? Beyond search and access in the NSDL. D-Lib Magazine, 11(11). Retrieved, January 10, 2007, from http://www.dlib.org/dlib/november05/lagoze/11lagoze.html.