GEOSCIENCE DATA CURATION USING A DIGITAL OBJECT MODEL AND OPEN-SOURCE FRAMEWORKS: PROVENANCE APPLICATIONS

Jerry Pan, Christopher Lenhardt, Bruce Wilson, Giri Palanisamy, Robert Cook, Biva Shrestha

Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6407, USA

1. INTRODUCTION

Scientific digital content, including Earth Science observations and model output, has become more heterogeneous in format and more distributed across the Internet. In addition, data and metadata are becoming necessarily linked internally and externally on the Web. As a result, such content has become more difficult for providers to manage and preserve and for users to locate, understand, and consume. Specifically, it is increasingly harder to deliver consistently relevant metadata and data processing lineage information along with the actual content. Readme files, data quality information, production provenance, and other descriptive metadata are often separated at the storage level as well as in the data search and retrieval interfaces available to a user. Critical archival metadata, such as auditing trails and integrity checks, are often even more difficult for users to access, if they exist at all. While these challenges exist for all science domains, in Geosciences they are compounded by large file sizes, multiple descriptive metadata standards (and evolving), heterogeneous file formats, and differing requirements of various data access/visualization tools. And yet, it is critically important to be able to store and disseminate such metadata in order for the data to be effectively preserved and reused for long term. In this study we explore a digital object model along with several open-source frameworks to solve this problem.

2. PROVENANCE FROM DATA ARCHIVE PERSPECTIVES

One important provenance issue is trust: how can the stakeholders verify authenticity, integrity, and reproducibility of a piece of data. Recent research and development in this area has shown great promise for the case where data is being created or transformed, particularly in a workflow environment. Standardized provenance annotation data model and schema - the Open Provenance Model (OPM) [6,8] has been designed and a few popular scientific workflow engines have started to support it. Among other things, OPM-compliant systems allow the capture of provenance and data lineage information when data is being created, as well as automated provenance information exchange between computer systems. Data can therefore be

reproduced at each transformation step from its inputs. For example, workflow engines like Kelper, Taverna, and eBioFlow all added OPM support [1], while the stand-alone, Karma Provenance Collection Tool [3] also has implemented the latest OPM.

However, less attention is currently being paid to the situation where datasets have already been created and ingested into various data archive centers: this is the after-the-fact case, which is different from the real time data creation case in that the provenance data may need to be manually collected from multiple sources. The majority of today's existing datasets fall into this category and a provenance-aware workflow solution like OPM may not apply. Given that not all science data creation are in a workflow setting, the situation of inadequate provenance capture will continue in the foreseeable future.

From the perspective of a data archive center, data provenance capture and storage is only part of the solution for an archive to be a full-fledged data curation system, because long-term preservation needs are more than simple data replication. A curation system needs to accommodate many different types of metadata, such as descriptive and preservation metadata, in additional to provenance metadata. A data curation system also needs to manage and identify the changes within the system (e.g., a data tool comes along and demands another format). Additionally, a curation system also needs to have user-oriented functionalities, such as rights control, user authentication and authorization, and data dissemination/visualization.

3. OPEN-SOURCE DIGITAL OBJECT FRAMEWORK: A VIABLE SOLUTION

We use a digital object abstraction framework as a data curation system to preserve and record provenance for the after-the-fact case mentioned above. The components of the abstraction can be used to capture provenance information such as creation history, descriptive metadata, code snippets or algorithms, or any arbitrary contextual information for the data. A number of open-source repository frameworks have a logic entity as the unit of management. For example, DSpace, Fedora Repository, and Storage Resource Broker all have logic management units that wraps on top of content (mostly files). Among these, only Fedora Repository provides a semantics framework to manage relationships among the components of a logical unit, as well as the relationships between logical units [4,10]. The relationships are expressed in Resource Description Framework (RDF) triples with a basic ontology, which is extensible by data practitioners, and the triples can be queried through related RDF technologies such as SPARQL. Traditionally, the relationships are used to map real world entities such as collections within datasets hierarchies. We tested whether these can also be used as a semantic store for provenance information for Geoscience datasets. The digital object encapsulates all relevant resources of digital content

and is itself encoded in formal xml records. The repository maybe rebuilt from these xml records alone.

With the digital object model, metadata of data description and data provenance can be associated with data content in a formal manner, so are external references and other arbitrary auxiliary information. Changes are formally audited on an object, and digital contents are versioned and have checksums automatically computed. Further, relationships among objects are formally expressed with RDF triples. Data replication, recovery, metadata export are supported with standard protocols, such as OAI-PMH [2]. The Fedora repository framework has all essential components as described in the Open Archival Information System (OAIS) reference model [11] and is compatible to that ISO standard.

Because the core of the Fedora Repository system is metadata agnostic (metadata and data are treated the same) and it supports multiple metadata schemas, datasets in Geoscience domain are readily supported with a Geocience metadata schema, such as FGDC-CSDGM or ISO19115-NAP. One of the main cost factors to use Fedora Repository for Geoscience data is the need to have an integrated metadata entry system for the selected metadata schema. The storage and discovery subsystems for any new metadata schemas are largely ready, through the core services provided by Fedora Repository. We use some of the built-in RDF statements to capture data derivation history, which is rendered as a graph view. In addition, we manage and describe data service endpoints as components of an object, which are related back to the data object via an RDF statement.

We also extend the basic Fedora ontology to include several custom relations, for the purpose to semantically mark a resource (file, web URL) as the metadata or the viewer for a particular data entity like a dataset. We construct the user interfaces with the semantic relations and other metadata records of both structured xml and plain text.

In addition to Fedora Repository, we use Drupal Content Management System (CMS) as the user-interface, the Islandora module [7] as the connector from Drupal to Fedora Repository, and Apache Solr as the search system.[12] These open source projects have significant user communities and enjoy continuous security scrutiny and code refactoring.

Our initial testing results using the terrestrial ecology data collections at NASA's ORNL Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC) [9] indicates that: (1) the digital object model works well as a logical store for both metadata and data, (2) the object semantics and built-in semantic store can be used to preserve provenance knowledge and other metadata, and (3) Drupal/Islandora UI is effective and extensible as the connector between the repository, the UI, and the searching system.

4. SUMMARY

We find that software frameworks for digital content management and access may be used for capturing certain data provenance information, particularly for data that has already been created and archived at a repository center. One of the key enabling factors is the abstraction concept of a digital object augmented with semantic relationships. One set of frameworks, Fedora Repository and Drupal CMS with the Islandora connector hold great promise for provenance applications as well as long-term curation of Geoscience datasets.

5. REFERENCES

[1] Altintas, O. Barney, and E. Jaeger-Frank, "Provenance collection support in the Kepler scientific workflow system," *Lecture Notes in Computer Science*, Volume 4145, Provenance and Annotation of Data, Pages 118-132, 2006.

[2] Bekaert, J. and H. Van de Sompel, "Access Interfaces for Open Archival Information Systems based on the OAI-PMH and the OpenURL Framework for Context-Sensitive Services," *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)*, The Royal Society: Edinburgh, UK., 2005.

[3] Cao, B., B. Plale, G. Subramanian, E. Robertson, and Y. Simmhan, "Provenance Information Model of Karma, IEEE," *Proc. Third International Workshop on Scientific Workflows*, Los Angeles, CA, 2009.

[4] fedora-commons.org. Fedora Commons. 2008, http://www.fedora-commons.org/.

[5] Freire, J., D. Koop, and L. Moreau, "Provenance and annotation of data and processes," *Lecture Notes in Computer Science*, Berlin: Springer. xi, 328 p., 2008.

[6] Futrelle, J., et al., "Semantic middleware for e-science knowledge spaces," *Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science, ACM*, Urbana Champaign, Illinois, 2009.

[7] Islandora Drupal Module, http://www.islandora.ca/.

[8] Moreau, L., I. Foster, and SpringerLink (Online service), "Provenance and annotation of data," *International Provenance and Annotation Workshop, IPAW 2006*, Springer: Berlin ; New York. p. xi, 288 p., 2006.

[9] ORNL DAAC: The Oak Ridge National Laboratory Distributed Active Archive Center, http://ornl.daac.gov/.

[10] Payette, S. and C. Lagoze, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA),". S*econd European Conference on Research and Advanced Technology for Digital Libraries*, Heraklion, Crete, Greece: Springer, 1998.

[11] Reference Model for an Open Archival Information System (OAIS): http://public.ccsds.org/publications/archive/650x0b1.pdf.

[12] Apache solr project, http://lucene.apache.org/solr/index.html.