NASA—Digital Object (ESDORA)

PRINCIPAL INVESTIGATOR: Jerry Y. Pan

PARTICIPATING STAFF: Giriprakash Palanisamy, Bruce E. Wilson, W. Christopher Lenhardt, Robert B. Cook, Biva Shrestha

PROJECT START DATE: January 10, 2010

PROJECT END DATE: December 31, 2011

SPONSOR: National Aeronautic and Space Administration (NASA)

PARTNERS: DuraSpace.org, University of Prince Edward Island, Johns Hopkins University

PROJECT WEBSITE: http://esdora.ornl.gov/

PROJECT DESCRIPTION

The goal for ESDORA is to deliver Earth sciences data, particularly the National Aeronautic and Space Administration's (NASA) Earth Observing System (EOS) data, using digital content repository technology, promoting best practice in digital provenance and enabling effective access to content and associated metadata.

Based on preliminary work, the ESDORA development team expects to couple the Fedora Repository and a Drupal-based graphic user interface (GUI) as key elements of a next-generation NASA Earth system science data center infrastructure, using datasets collected as part of the North American Carbon Program (NACP) Modeling and Synthesis Thematic Data Center (MAST-DC) as the primary science context. The team will use this implementation to enable better and more consistent access to critical metadata, including processing lineage information and administrative metadata, using the capabilities inherent in a digital repository (multiple streams for a given object and remote data streams). ESDORA seeks to demonstrate how data providers can more easily and effectively manage science data sets, associated metadata, processing lineage, and quality control/data provenance information. A consistent process, with associated user interfaces and application programming interfaces (APIs), can be used by the data provider to ingest, update, and modify a dataset for metadata changes or additional content dissemination revenues. Particularly in the context of the ORNL Distributed Active Archive Center data, this work will demonstrate potential technology migration paths for existing data operations.

SIGNIFICANCE

Digital content, including Earth science observations and model output, is an essential part of contemporary scientific research activities. Not only is the rate of archiving for such content increasing rapidly, but there is also an increase in derived and on-demand data product creation and consumption. As a result of these trends, scientific digital content has become even more heterogeneous in format and more distributed across the Internet. In turn, this makes the content more difficult for providers to manage and preserve and for users to locate, understand, and consume. Specifically, it is increasingly harder to deliver relevant metadata and data processing lineage information along with the actual content, particularly when there are multiple ways of delivering the content, including the increasing use of web services. Readme files, data quality information, production provenance, and other descriptive metadata are often separated in the storage level as well 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. The development team behind ESDORA proposes to address these challenges by using and extending the capabilities of a contemporary digital object repository to work for science data and metadata delivery. Digital repository technology has been used for digital libraries with great success, and ESDORA seeks to apply this technology to the more complex needs of Earth science data management.

INTERESTING FINDINGS

Application of digital repository technology to data management is a new area of exploration, and this project is a pioneering experiment in this area. In our development effort, some of the advertised benefits as described in the above sections are largely attested and confirmed. For example, in metadata management, we have successfully configured

the system to use the FGDC-CSDGM (Federal Geographic Data Committee, Content Standard for Digital Geospatial Metadata) metadata record in addition to the required DC (Dublin Core) metadata for a data collection, from which chosen fields in both metadata records are searchable in the system. It should also be the case that an existing repository can change or add the metadata with a new standard such as ISO 19115. Additionally, dataset identification using Digital Object Identifiers (DOIs) is readily enabled by this system. The system is flexible and customizable, and we have just completed a modification to allow the file storage of content to be based on human-understandable hierarchy in order for the content to be easily accessible from other tools. However, there are also some challenges in using and promoting the technology. One such area, how to model content into digital objects (content modeling), is quite subjective, and it can vary from dataset to dataset. In addition, once a content model is used for real data, it is quite rigid, and radical changes of the model may prove to be costly for production system. The management and user access module (Islandora) is undergoing a lot of changes toward a more positive and user-friendly interface, but compared to mature, open source software, there are still some gaps in community adoption, documentation, and overall maturity.