	Control ID	Final ID	Title	Presenting Author
1	1804484		Embedding Data Stewardship in Geoscience Australia	Bastrakova, Irina
2	1820232		Facilitating Stewardship of scientific data through standards based workflows	Bastrakova, Irina
3	1817159		Outcomes of the "Data Curation for Geobiology at Yellowstone National Park" Workshop	Thomer, Andrea
4	1817468		Rolling Deck to Repository (R2R): Supporting Global Data Access Through the Ocean Data Interoperability Platform (ODIP)	Arko, Robert
5	1820307		Provenance Capture in Data Access and Data Manipulation Software	West, Patrick
6	1818936		Ensuring the Quality of Data Packages in the LTER Network Provenance Aware Synthesis Tracking Architecture Data Manage…	O'Brien, Margaret
7	1821000		Between land and sea: divergent data stewardship practices in deep-sea biosphere research	Cummings, Rebekah

8	1808246	ESIP's Emerging Provenance and Context Content Standard Use Cases: Developing Examples and Models for Data Stewardship	Ramdeen, Sarah
9	1803280	Dataset Lifecycle Policy Development & Implementation at the PO.DAAC	Moroni, David
10	1816771	'Best' Practices for Aggregating Subset Results from Archived Datasets	Baskin, Walter
11	1815474	Don't leave data unattended at any time!	Fleischer, Dirk

Print Close Window

Proof

CONTROL ID: 1804484

TITLE: Embedding Data Stewardship in Geoscience Australia

PRESENTATION TYPE: Poster Requested

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Irina Bastrakova[1], Sue Fyfe[1]

INSTITUTIONS (ALL): 1. Geoscience Australia, Canberra, ACT, Australia.

PRESENTER: Irina Bastrakova

PRESENTER (E-MAIL ONLY): irina.bastrakova@ga.gov.au

ABSTRACT BODY: Ten years of technological innovation now enable vast amounts of data to be collected, managed, processed and shared. At the same time, organisations have witnessed government legislative and policy requirements for open access to public sector data, and a demand for flexibility in access to data by both machine-to-machine and human consumption.

Geoscience Australia (GA) has adopted Data Stewardship as an organisation-wide initiative to improve the way we manage and share our data. The benefits to GA including:

- Consolidated understanding of GA's data assets and their value to the Agency;
- Recognition of the significant role of data custodianship and data management;
- Well-defined governance, policies, standards, practices and accountabilities that promote the accessibility, quality and interoperability of GA's data;
- Integration of disparate data sets into cohesive information products available online in real time and equally accessible to researchers, government, industry and the public.

Although the theory behind data stewardship is well-defined and accepted and the benefits are generally well-understood, practical implementation requires an organisation to prepare for a long-term commitment of resources, both financial and human. Fundamentally this involves:

1. Raising awareness in the organisation of the need for data stewardship and the challenges this entails;
2. Establishing a data stewardship framework including a data governance office to set policy and drive organisational change; and
3. Embedding the functions and a culture of data stewardship into business as usual operations.

GA holds a vast amount of data ranging from petabytes of Big Data to significant quantities of relatively small 'long tail' geoscientific observations and measurements. Over the past four years, GA has undertaken strategic activities that prepare us for Data Stewardship:

- Organisation-wide audits of GA's data holdings and identification of custodians for each dataset;
- Developing guiding Principles on how the Agency undertakes Science, Data Management and Cataloguing;
- Developing a Data Classification Schema that aligns scientific requirements and business workflows with data architecture;
- Creating the Scientific Data Stewardship Steering Committee, comprising champions from across the Agency to guide development and support implementation of Data Stewardship in GA;
- Forming Scientific Data Communities of Practice of leading scientific experts to identify tandards and

practices across their domain, and integrate data stewardship practices into scientific workflows;
- Establishing the Data Governance and Services Section to provide ongoing capacity for the development, communication, coordination and governance of data stewardship policies, strategies, standards and practices.

GA is now moving towards Data Stewardship as an operational capability and culture within the Agency. The challenges we face into the future include:

- Maintaining continuous and enthusiastic engagement from the Agency executive;
- Implementing long term cultural change at all levels within the organisation;
- Formal recognition that data stewardship is a continuous operational BAU activity;
- Incorporation of data custodianship and management activities in work programs and budgets; and
- Cultivation and support of the data stewardship champions.

INDEX TERMS: 1912 INFORMATICS Data management, preservation, rescue, 1930 INFORMATICS Data and information governance, 1916 INFORMATICS Data and information discovery, 1938 INFORMATICS Knowledge representation and knowledge bases.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1820232

TITLE: Facilitating Stewardship of scientific data through standards based workflows

PRESENTATION TYPE: Poster Requested

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Irina Bastrakova[1], Carina Kemp[1], Anna Kathryn Potter[1]

INSTITUTIONS (ALL): 1. Geoscience Australia, Canberra, ACT, Australia.

PRESENTER: Irina Bastrakova

PRESENTER (E-MAIL ONLY): irina.bastrakova@ga.gov.au

ABSTRACT BODY: There are main suites of standards that can be used to define the fundamental scientific methodology of data, methods and results. These are firstly Metadata standards to enable discovery of the data (ISO 19115), secondly the Sensor Web Enablement (SWE) suite of standards that include the O&M and SensorML standards and thirdly Ontology that provide vocabularies to define the scientific concepts and relationships between these concepts. All three types of standards have to be utilised by the practicing scientist to ensure that those who ultimately have to steward the data stewards to ensure that the data can be preserved curated and reused and repurposed. Additional benefits of this approach include transparency of scientific processes from the data acquisition to creation of scientific concepts and models, and provision of context to inform data use.

Collecting and recording metadata is the first step in scientific data flow. The primary role of metadata is to provide details of geographic extent, availability and high-level description of data suitable for its initial discovery through common search engines.

The SWE suite provides standardised patterns to describe observations and measurements taken for these data, capture detailed information about observation or analytical methods, used instruments and define quality determinations. This information standardises browsing capability over discrete data types. The standardised patterns of the SWE standards simplify aggregation of observation and measurement data enabling scientists to transfer disintegrated data to scientific concepts.

The first two steps provide a necessary basis for the reasoning about concepts of ‘pure’ science, building relationship between concepts of different domains (linked-data), and identifying domain classification and vocabularies.

Geoscience Australia is re-examining its marine data flows, including metadata requirements and business processes, to achieve a clearer link between scientific data acquisition and analysis requirements and effective interoperable data management and delivery. This includes participating in national and international dialogue on development of standards, embedding data management activities in business processes, and developing scientific staff as effective data stewards.

Similar approach is applied to the geophysical data. By ensuring the geophysical datasets at GA strictly follow metadata and industry standards we are able to implement a provenance based workflow where the data is easily discoverable, geophysical processing can be applied to it and results can be stored. The provenance based workflow enables metadata records for the results to be produced automatically from the input dataset metadata.

INDEX TERMS: 1946 INFORMATICS Metadata, 1936 INFORMATICS Interoperability.
(No Table Selected)
(No Image Selected)

Additional Details
Previously Presented Material:
Scheduling Request:
TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1817159

TITLE: Outcomes of the "Data Curation for Geobiology at Yellowstone National Park" Workshop

PRESENTATION TYPE: Poster Requested

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Andrea Thomer[1], Carole L. Palmer[1], Bruce W Fouke[2], Ann Rodman[3], G. Sayeed Choudhury[4], Karen S. Baker[1], Abigail Esenam Asangba[2], Karen Wickett[1], Timothy DiLauro[4], Virgil Varvel[1]

INSTITUTIONS (ALL): 1. Center for Informatics Research in Science and Scholarship, University of Illinois at Urbana-Champaign, Champaign, IL, United States.
2. Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL, United States.
3. Resource Information Management, Geographic Information Systems,, Yellowstone National Park, Gardiner, MT, United States.
4. Digital Research and Curation Center at the Sheridan Libraries, Johns Hopkins University, California, MD, United States.

PRESENTER: Andrea Thomer

PRESENTER (E-MAIL ONLY): thomer2@illinois.edu

ABSTRACT BODY: The continuing proliferation of geological and biological data generated at scientifically significant sites (such as hot springs, coral reefs, volcanic fields and other unique, data-rich locales) has created a clear need for the curation and active management of these data. However, there has been little exploration of what these curation processes and policies would entail. To that end, the Site-Based Data Curation (SBDC) project is developing a framework of guidelines and processes for the curation of research data generated at scientifically significant sites. A workshop was held in April 2013 at Yellowstone National Park (YNP) to gather input from scientists and stakeholders. Workshop participants included nine researchers actively conducting geobiology research at YNP, and seven YNP representatives, including permitting staff and information professionals from the YNP research library and archive. Researchers came from a range of research areas -- geology, molecular and microbial biology, ecology, environmental engineering, and science education.

Through group discussions, breakout sessions and hands-on activities, we sought to generate policy recommendations and curation guidelines for the collection, representation, sharing and quality control of geobiological datasets. We report on key themes that emerged from workshop discussions, including:

- participants' broad conceptions of the long-term usefulness, reusability and value of data.
- the benefits of aggregating site-specific data in general, and geobiological data in particular.
- the importance of capturing a dataset's originating context, and the potential usefulness of photographs as a reliable and easy way of documenting context.
- researchers' and resource managers' overlapping priorities with regards to "big picture" data collection and management in the long-term.

Overall, we found that workshop participants were enthusiastic and optimistic about future collaboration and development of community approaches to data sharing. We hope to continue discussion of geobiology data curation challenges and potential strategies at AGU. Outcomes from the workshop are guiding next steps in

the SBDC project, led by investigators at the Center for Informatics Research in Science and Scholarship and Institute for Genomic Biology at the University of Illinois, in collaboration with partners at Johns Hopkins University and YNP.

INDEX TERMS: 1912 INFORMATICS Data management, preservation, rescue, 1946 INFORMATICS Metadata, 1999 INFORMATICS General or miscellaneous, 1904 INFORMATICS Community standards.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1817468

TITLE: Rolling Deck to Repository (R2R): Supporting Global Data Access Through the Ocean Data Interoperability Platform (ODIP)

PRESENTATION TYPE: Poster Requested

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Robert A Arko[1], Karen Stocks[2], Cynthia L Chandler[3], Shawn R Smith[4], Stephen P Miller[2], Andrew Richard Maffei[3], Helen M Glaves[5], Suzanne M Carbotte[1]

INSTITUTIONS (ALL): 1. Lamont-Doherty Earth Observatory, Palisades, NY, United States.
2. Scripps Institution of Oceanography, La Jolla, CA, United States.
3. Woods Hole Oceanographic Institution, Woods Hole, MA, United States.
4. Florida State University, Tallahassee, FL, United States.
5. British Geological Survey, Northampton, United Kingdom.

PRESENTER: Robert Arko

PRESENTER (E-MAIL ONLY): arko@ldeo.columbia.edu

ABSTRACT BODY: The U.S. National Science Foundation supports a fleet of academic research vessels operating throughout the world's oceans. In addition to supporting the mission-specific goals of each expedition, these vessels routinely deploy a suite of underway environmental sensors, operating like mobile observatories. Recognizing that the data from these instruments have value beyond each cruise, NSF funded R2R in 2009 to ensure that these data are routinely captured, cataloged and described, and submitted to the appropriate national repository for long-term public access.

In 2013, R2R joined the Ocean Data Interoperability Platform (ODIP; http://odip.org/). The goal of ODIP is to remove barriers to the effective sharing of data across scientific domains and international boundaries, by providing a forum to harmonize diverse regional systems. To advance this goal, ODIP organizes international workshops to foster the development of common standards and develop prototypes to evaluate and test potential standards and interoperability solutions. ODIP includes major organizations engaged in ocean data stewardship in the EU, US, and Australia, supported by the International Oceanographic Data and Information Exchange (IODE).

Within the broad scope of ODIP, R2R focuses on contributions in 4 key areas:
● Implement a "Linked Open Data" approach to disseminate data and documentation, using existing World Wide Web Consortium (W3C) specifications and machine-readable formats. Exposing content as Linked Open Data will provide a simple mechanism for ODIP collaborators to browse and compare data sets among repositories.
● Map key vocabularies used by R2R to their European and Australian counterparts. The existing heterogeneity among terms inhibits data discoverability, as a user searching on the term with which s/he is familiar may not find all data of interest. Mapping key terms across the different ODIP partners, relying on the backbone thesaurus provided by the NERC Vocabulary Server (http://vocab.nerc.ac.uk/), is a first step towards wider data discoverability.
● Upgrade existing R2R ISO metadata records to be compatible with the new SeaDataNet II Cruise Summary Report (CSR) profile, and publish the records in a standards-compliant Web portal, built on the GeoNetwork open-source package.

● Develop the future workforce. R2R is enlisting and exposing a group of five students to new informatics technologies and international collaboration. Students are undertaking a coordinated series of projects in 2013 and 2014 at each of the R2R partner institutions, combined with travel to selected meetings where they will engage in the ODIP Workshop process; present results; and exchange ideas with working scientists and software developers in Europe and Australia. Students work closely with staff at the R2R partner institutions, in projects that build on the R2R-ODIP technical work components described above.

http://rvdata.us/

INDEX TERMS: 1934 INFORMATICS International collaboration, 1936 INFORMATICS Interoperability, 1916 INFORMATICS Data and information discovery, 4200 OCEANOGRAPHY: GENERAL.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1820307

TITLE: Provenance Capture in Data Access and Data Manipulation Software

PRESENTATION TYPE: Poster Requested

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Patrick West[1], Peter Arthur Fox[1], Deborah L McGuinness[1], James H R Gallagher[2], Dan Holloway[2], Nathan Potter[2]

INSTITUTIONS (ALL): 1. Tetherless World Constellation, Rensselaer Polytechnic Institu, Troy, NY, United States.
2. OPeNDAP, Narragansett, RI, United States.

PRESENTER: Patrick West

PRESENTER (E-MAIL ONLY): westp@rpi.edu

ABSTRACT BODY: There is increasing need to trace back the origins of data products, whether images or charts in a report, data obtained from a sensor on an instrument, a generated dataset referenced in a research paper, in government reports on the environment, or in a publication or poster presentation. Yet, most software applications that perform data access and manipulation keep only limited history of the data, i.e. the provenance. Imagine the following scenario: There is a figure in a report showing multiple graphs and plots related to global climate, the report is being drafted for a government agency. The graphs and plots are generated using an algorithm from an iPython Notebook, developed by a researcher who is using a particular data portal, where the algorithm pulls data from four data sets from that portal. That data is aggregated together over the time dimension, constrained to a few parameters, accessed using a particular piece of data access software, and converted from one datatype to another datatype; All the processing on the data sets was conducted by three different researchers from a public university, on a project funded by the same government agency requesting the report, with one Principal Investigator and two Co-Investigators. In this scenario, today we're lucky to get a blob of text under the figure that might say a couple things about the figure with a reference to a publication that was written a few years ago. Data citation, data publishing information, licensing information, and provenance are all lacking in the derived data products.

What we really want is to be able to trace the figure all the way back to the original datasets, including what was done to those datasets; and to see information about the researchers, the project, the agency funding, the award, and the organizations collaborating on the project. In this paper we discuss the need for such information and traceback features, as well as new technologies and standards that can help us become better data stewards. Specifically, we will talk about the new PROV recommendation from the W3C, recently published, and existing and new features in the OPeNDAP software stack that can help facilitate the incorporation of citation, licensing, and provenance information and the ability to click through to retrieve that information.

INDEX TERMS: 1916 INFORMATICS Data and information discovery, 1970 INFORMATICS Semantic web and semantic integration, 1938 INFORMATICS Knowledge representation and knowledge bases, 1976 INFORMATICS Software tools and services.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1818936

TITLE: Ensuring the Quality of Data Packages in the LTER Network Provenance Aware Synthesis Tracking Architecture Data Management System and Archive

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Mark Stephen Servilla[1], Margaret O'Brien[2], Duane Costa[1]

INSTITUTIONS (ALL): 1. LTER Network Office, University of New Mexico, Albuquerque, NM, United States.
2. Santa Barbara Coastal LTER, University of California, Santa Barbara, Santa Barbara, CA, United States.

PRESENTER: Margaret O'Brien

PRESENTER (E-MAIL ONLY): mob@msi.ucsb.edu

ABSTRACT BODY: Considerable ecological research performed today occurs through the analysis of data downloaded from various repositories and archives, often resulting in derived or synthetic products generated by automated workflows. These data are only meaningful for research if they are well documented by metadata, lest semantic or data type errors may occur in interpretation or processing. The Long Term Ecological Research (LTER) Network now screens all data packages entering its long-term archive to ensure that each package contains metadata that is complete, of high quality, and accurately describes the structure of its associated data entity and the data are structurally congruent to the metadata. Screening occurs prior to the upload of a data package into the Provenance Aware Synthesis Tracking Architecture (PASTA) data management system through a series of quality checks, thus preventing ambiguously or incorrectly documented data packages from entering the system. The quality checks within PASTA are designed to work specifically with the Ecological Metadata Language (EML), the metadata standard adopted by the LTER Network to describe data generated by their 26 research sites. Each quality check is codified in Java as part of the ecological community-supported Data Manager Library, which is a resource of the EML specification and used as a component of the PASTA software stack. Quality checks test for metadata quality, data integrity, or metadata-data congruence. Quality checks are further classified as either conditional or informational. Conditional checks issue a “valid”, “warning” or “error” response. Only an “error” response blocks the data package from upload into PASTA. Informational checks only provide descriptive content pertaining to a particular facet of the data package. Quality checks are designed by a group of LTER information managers and reviewed by the LTER community before deploying into PASTA. A total of 32 quality checks have been deployed to date. Quality checks can be customized through a configurable template, which includes turning checks “on” or “off” and setting the severity of conditional checks. This feature is important to other potential users of the Data Manager Library who wish to configure its quality checks in accordance with the standards of their community. Executing the complete set of quality checks produces a report that describes the result of each check. The report is an XML document that is stored by PASTA for future reference.

INDEX TERMS: 1900 INFORMATICS, 1920 INFORMATICS Emerging informatics technologies, 1950 INFORMATICS Metadata: Quality.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1821000

TITLE: Between land and sea: divergent data stewardship practices in deep-sea biosphere research

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Rebekah Cummings[1], Peter Darch[1]

INSTITUTIONS (ALL): 1. University of California Los Angeles, Salt Lake City, UT, United States.

PRESENTER: Rebekah Cummings

PRESENTER (E-MAIL ONLY): rebekahlynncummings@gmail.com

ABSTRACT BODY: Data in deep-sea biosphere research often live a double life. While the original data generated on IODP expeditions are highly structured, professionally curated, and widely shared, the downstream data practices of deep-sea biosphere laboratories are far more localized and ad hoc. These divergent data practices make it difficult to track the provenance of datasets from the cruise ships to the laboratory or to integrate IODP data with laboratory data. An in-depth study of the divergent data practices in deep-sea biosphere research allows us to:

- Better understand the social and technical forces that shape data stewardship throughout the data lifecycle;
- Develop policy, infrastructure, and best practices to improve data stewardship in small labs;
- Track provenance of datasets from IODP cruises to labs and publications;
- Create linkages between laboratory findings, cruise data, and IODP samples.

In this paper, we present findings from the first year of a case study of the Center for Dark Energy Biosphere Investigations (C-DEBI), an NSF Science and Technology Center that studies life beneath the seafloor. Our methods include observation in laboratories, interviews, document analysis, and participation in scientific meetings. Our research uncovers the data stewardship norms of geologists, biologists, chemists, and hydrologists conducting multi-disciplinary research.

Our research team found that data stewardship on cruises is a clearly defined task performed by an IODP curator, while downstream it is a distributed task that develops in response to local need and to the extent necessary for the immediate research team.

IODP data are expensive to collect and challenging to obtain, often costing $50,000/day and requiring researchers to work twelve hours a day onboard the ships. To maximize this research investment, a highly trained IODP data curator controls data stewardship on the cruise and applies best practices such as standardized formats, proper labeling, and centralized storage.

In the laboratory, a scientist is his or her own curator. In contrast to the IODP research parties, laboratory research teams analyze diverse datasets, share them internally, implement ad hoc data management practices, optimize methods for their specific research questions, and release data on request through personal transactions. We discovered that while these workflows help small research teams retain flexibility and local control -- crucial in exploratory deep-sea biosphere research -- they also hinder data interoperability, discoverability, and consistency of methods from one research team to the next.

Additional consequences of this contrast between IODP and lab practices are that it is difficult to track the provenance of data and to create linkages between laboratory findings, cruise data, and archived IODP

samples. The ability to track provenance would add value to datasets and provide a clearer picture of the decisions made throughout the data lifecycle. Better linkages between the original data, laboratory data, and samples would allow secondary researchers to locate IODP data that may be useful to their research after laboratory findings are published.

Our case study is funded by the Sloan Foundation and NSF.

INDEX TERMS: 1916 INFORMATICS Data and information discovery.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM: UCLA Knowledge Infrastructures Team

Print Close Window

Proof

CONTROL ID: 1808246

TITLE: ESIP's Emerging Provenance and Context Content Standard Use Cases: Developing Examples and Models for Data Stewardship

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Sarah Ramdeen[1], Denise J Hills[2]

INSTITUTIONS (ALL): 1. University of North Carolina at Chapel Hill, Chapel Hill, NC, United States.
2. Geological Survey of Alabama, Tuscaloosa, AL, United States.

PRESENTER: Sarah Ramdeen

PRESENTER (E-MAIL ONLY): ramdeen@email.unc.edu

ABSTRACT BODY: Earth science data collections range from individual researchers' private collections to large-scale data warehouses, from computer-generated data to field or lab based observations. These collections require stewardship. Fundamentally, stewardship ensures long term preservation and the provision of access to the user community. In particular, stewardship includes capturing appropriate metadata and documentation--and thus the context of the data's creation and any changes they underwent over time --to enable data reuse. But scientists and science data managers must translate these ideas into practice. How does one balance the needs of current and (projected) future stakeholders?

In 2011, the Data Stewardship Committee (DSC) of the Federation of Earth Science Information Partners (ESIP) began developing the Provenance and Context Content Standard (PCCS). As an emerging standard, PCCS provides a framework for 'what' must be captured or preserved as opposed to describing only 'how' it should be done. Originally based on the experiences of NASA and NOAA researchers within ESIP, the standard currently provides data managers with content items aligned to eight key categories. While the categories and content items are based on data life cycles of remote sensing missions, they can be generalized to cover a broader set of activities, for example, preservation of physical objects. These categories will include the information needed to ensure the long-term understandability and usability of earth science data products.

In addition to the PCCS, the DSC is developing a series of use cases based on the perspectives of the data archiver, data user, and the data consumer that will connect theory and practice. These cases will act as specifications for developing PCCS-based systems. They will also provide for examination of the categories and content items covered in the PCCS to determine if any additions are needed to cover the various use cases, and also provide rationale and indicate priorities for preservation. Though the use cases currently focus on two areas, 'creating' a data set and 'using' a data set, the use cases will eventually cover the full data lifecycle. Currently developing a template to be used in future use case creation, the DSC is also preparing and testing more use case scenarios.

This presentation will introduce the ESIP use cases based on the PCCS. It will at once expand stakeholder participation and show the application of these materials beyond the ESIP community in which they were developed. More information about the ESIP use case activities can be found on the DSC wiki - http://wiki.esipfed.org/index.php/Preservation_Use_Case_Activity.

http://wiki.esipfed.org/index.php/Preservation_Use_Case_Activity

INDEX TERMS: 1900 INFORMATICS, 1948 INFORMATICS Metadata: Provenance, 1930 INFORMATICS Data and information governance, 1904 INFORMATICS Community standards.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM: Data Stewardship Committee for the Federation of Earth Science Information Partners

Print Close Window

Proof

CONTROL ID: 1803280

TITLE: Dataset Lifecycle Policy Development & Implementation at the PO.DAAC

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Vardis M Tsontos[1], Eric Tauer[1], David F Moroni[1], Edward M Armstrong[1], Jessica Hausman[1]

INSTITUTIONS (ALL): 1. PO.DAAC, NASA/JPL, Pasadena, CA, United States.

PRESENTER: David Moroni

PRESENTER (E-MAIL ONLY): David.F.Moroni@jpl.nasa.gov

ABSTRACT BODY: As part of on-going efforts to advance data stewardship standards and promote best practices in the field of earth science data management, NASA's Physical Oceanography Data archive (PO.DAAC), has developed and implemented a dataset lifecycle policy that governs the curation of our satellite data holdings. This lifecycle policy provides a framework, formalized process and requirements checklist useful for planning and managing datasets through the various phases of data archival at the PO.DAAC. Here we provide an overview of this framework including its motivation and the various lifecycle stages identified, from dataset identification/prioritization to archival planning, integration/testing, operations, ultimately through to retirement. Each phase is described in terms of its purpose, policy, associated artifacts and exit criteria, and the actors and roles involved. The generality of the framework to handle version control scenarios and the adaptability of the approach is reviewed. Usage of the dataset lifecycle policy as a tool for characterizing the status of PO.DAAC datasets and as metric that feeds into assessment of dataset maturity are also discussed.

INDEX TERMS: 1912 INFORMATICS Data management, preservation, rescue, 1998 INFORMATICS Workflow, 1982 INFORMATICS Standards, 1900 INFORMATICS.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1816771

TITLE: 'Best' Practices for Aggregating Subset Results from Archived Datasets

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Walter E Baskin[1], Jennifer Perez[2]

INSTITUTIONS (ALL): 1. NASA ASDC (SSAI Contractor), Hampton, VA, United States.
2. ASDC, NASA Langley Research Center, Hampton, VA, United States.

PRESENTER: Walter Baskin

PRESENTER (E-MAIL ONLY): Walter.E.Baskin@nasa.gov

ABSTRACT BODY: In response to the exponential growth in science data analysis and visualization capabilities Data Centers have been developing new delivery mechanisms to package and deliver large volumes of aggregated subsets of archived data. New standards are evolving to help data providers and application programmers deal with growing needs of the science community. These standards evolve from the best practices gleaned from new products and capabilities.

The NASA Atmospheric Sciences Data Center (ASDC) has developed and deployed production provider-specific search and subset web applications for the CALIPSO, CERES, TES, and MOPITT missions.

This presentation explores several use cases that leverage aggregated subset results and examines the standards and formats ASDC developers applied to the delivered files as well as the implementation strategies for subsetting and processing the aggregated products. The following topics will be addressed:
- Applications of NetCDF CF conventions to aggregated level 2 satellite subsets
- Data-Provider-Specific format requirements vs. generalized standards
- Organization of the file structure of aggregated NetCDF subset output
- Global Attributes of individual subsetted files vs. aggregated results
- Specific applications and framework used for subsetting and delivering derivative data files

INDEX TERMS: 1904 INFORMATICS Community standards, 1912 INFORMATICS Data management, preservation, rescue, 1910 INFORMATICS Data assimilation, integration and fusion, 1982 INFORMATICS Standards.

(No Table Selected)

(No Image Selected)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM:

Print Close Window

Proof

CONTROL ID: 1815474

TITLE: Don't leave data unattended at any time!

PRESENTATION TYPE: Assigned by Committee (Oral or Poster)

CURRENT SECTION/FOCUS GROUP: Earth and Space Science Informatics (IN)

CURRENT SESSION: IN011. Data Stewardship: in Theory and in Practice

AUTHORS (FIRST NAME, LAST NAME): Dirk Fleischer[1], Andreas Czerniak[1], Carsten Schirnick[1]

INSTITUTIONS (ALL): 1. GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.

PRESENTER: Dirk Fleischer

PRESENTER (E-MAIL ONLY): dfleischer@geomar.de

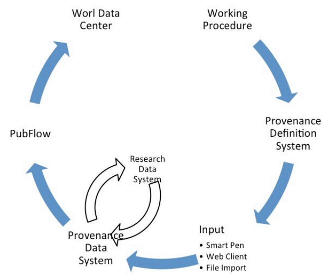
ABSTRACT BODY: The architecture of Kiel Data Management Infrastructure (KDMI) is setup to serve from the data creation process all the way to the data publication procedure. Accordingly the KDMI is managing data at the right beginning of the data life cycle and does not leave data unattended at this very crucial time. Starting from the chosen working procedure to handwritten protocols or lab notes the provenance of the resulting research data is captured within the KDMI.
The provenance definition system is the fundamental (see figure 1) capturing tool for working procedures. The provenance definition is used to enable data input by file import, web client or hand writing recognition. The captured data in the provenance system for data is taking care of unpublished in house research data created directly on site. This system serves as a master for research data systems with more degrees of freedom in regard to technology, design or performance (e.g. GraphDB, etc). Such research systems can be regarded as compilations of unpublished data and public domain data e.g. from World Data Centers or archives. These compilations can be used to run statistical data mining and pattern finding algorithms on these specially designed platforms.
The architecture of the KDMI ensures that a technical solution for data correction from the slave systems to the master system is possible and improves the quality of the stored data in the provenance system for data. After the research phase is over and the interpretation is finished the provenance system is used by a workflow based publication system called PubFlow. Within PubFlow it is possible to create repeatable workflows to publish data into various external long-term archives or World Data Center. The KDMI is based on the utilization of persistent identifiers for samples and person identities to support this automatized publication process. The publication process is the final step of the KDMI and the management responsibility of the long-term part of the data life cycle is handed over to the chosen archive. Nevertheless the provenance information remains at the KDMI and the definition maybe serves for future datasets again.

Unattended data may get lost or be destroyed

http://portal.geomar.de

INDEX TERMS: 1900 INFORMATICS, 1912 INFORMATICS Data management, preservation, rescue.

(No Table Selected)



Usage cycle of the Kiel Data Management Infrastructure (KDMI)

Additional Details

Previously Presented Material:

Scheduling Request:

TITLE OF TEAM: Kiel Data Management Team