

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

AGU
Poster #IN53C-1578

San Francisco, CA December 13th, 2013

Data Stewardship Committee for the Federation of
Earth Science Information Partners

Sarah Ramdeen ramdeen@email.unc.edu
SILS University of North Carolina at Chapel Hill

Denise J Hills dhills@gsa.state.al.us
Geological Survey of Alabama



Background

Provenance and Context Content Standard

PCCS provides a framework for 'what' must be captured or preserved as opposed to describing only 'how' it should be done. The standard currently provides data managers with content items aligned to eight key categories:

1. Preflight/Pre-Operations: Instrument/Sensor characteristics including pre-flight/pre-operations performance measurements; calibration method; radiometric and spectral response; noise characteristics; detector offsets
2. Products (Data): Raw instrument data, Level 0 through Level 4 data products and associated metadata
3. Product Documentation: Structure and format with definitions of all parameters and metadata fields; algorithm theoretical basis; processing history and product version history; quality assessment information
4. Mission Calibration: Instrument/sensor calibration method (in operation) and data; calibration software used to generate lookup tables; instrument and platform events and maneuvers
5. Product Software: Product generation software and software documentation
6. Algorithm Input: Any ancillary data or other data sets used in generation or calibration of the data or derived product; ancillary data description and documentation
7. Validation: Record and data sets
8. Software Tools: product access (reader) tools.

http://www.nasa.gov/images/content/291999main_quikscat-large.jpg



1. Summary

Definition: Brief description of the use case, addressing who is the intended audience and/or key reasons why someone may be interested in the example.

Example:

A research user needs to pick the data set from multiple similar data sets that best meets the user's requirements for their intended application. An example could be a polar bear ecologist choosing a data set on sea ice conditions in a region of the Hudson Bay from the multiple data sets listed at NSIDC. Another example could be a user choosing which sea surface temperature data set from PO.DAAC to use in forcing a model of an algal bloom. Many other examples exist. Traditionally this was done by the user consulting a relevant expert. Ideally, one could conceive of an expert system helping guide the user through their query, if the system had access to sufficient information.

Use case template

2. Objective/Context

Definition: Goal or intended focus of the use case, more specific and narrow than the summary.

Example:

Choosing a data set from multiple similar choices.



http://www.nasa.gov/images/content/60197main_remote_sensing_pigeons.jpg

3. Actors

Definition: Will vary depending on the scenario. Usually those who have a role in the use case. This includes relevant experts, possible contributors, or users of the data specified in the case.

Example:

Research user
Data expert(s)/Expert system:
• **science domain experts** that know the science applications.
• **instrument experts** that know the subtleties of the observation mechanism.
• **algorithm experts** that know the variations in retrievals.
• **process experts** that know the subtleties of the processing implementation.
• **data format experts** that know handling of for example HDF4 vs. HDF5.
Archive

4. Sequence of Events

Definition: Steps, in order of occurrence (where possible) detailing the actions and events of the use case.

Example:

1. User poses initial request to expert
2. Expert queries user on specifics
3. Iteration between user and expert to understand vocabularies and actual needs
 - a. Initial possible data sets are identified by basic criteria like whether the data set covers the right time and location
 - b. The list is further refined by more qualitative criteria specific to the actual query
4. A recommended data set or ranked list of data sets is returned to the user

Use Case Categories

Creating a data set:

- Releasing a data set*
- Publishing a data set *
- Creating a long term trend data set from multiple data sets*
 - A researcher 100 years in the future is examining the historical record.
- Giving credit to people involved in the data set
- Describing appropriate uses of a data set
 - Asserting quality of data set
 - Intellectual property rights
 - Policies
 - Creating citations for published data set
 - Validation of data

Using a data set:

- Obtaining Data*
 - Discovering data*
 - Preliminary search*
 - Choosing a data set from multiple similar choices*
 - Assessing data
- Citing use of a data set
- Applications for data
 - Analysis of new version of data set
 - How do the improvements affect our use of the data?
 - Sharing data set for collaboration
 - Comparing multiple data sets
 - Reproducing a dataset*
 - Verification of an experiment

* Indicates an existing case study

Acknowledgements

The authors would like to thank everyone in the ESIP community who has contributed to the development of the PCCS, the use case activities and this project.

For more information or to contribute to the use case activity process, please see our website:
http://wiki.esipfed.org/index.php/Preservation_Use_Case_Activity.

Benefits of use cases: creators and users

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). It is based on the experiences of NASA and NOAA researchers within ESIP. The categories and content items are based on data life cycles of remote sensing missions, however they can be generalized to cover a broader set of activities, for example, preservation of physical objects. These categories include the information needed to ensure the long-term understandability and usability of earth science data products.

Our 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. Our use cases demonstrate the perspectives of the data archivist, data user, and the data consumer as well as the balance between theory and practice.

Use case activities

- Identify what content items are required and why
- Mapping of use cases to content items helps with rationale and priority for preserving them
- If no use cases are identified for some of the items
 - Look harder for use cases, or
 - Eliminate item from PCCS

Keep in Mind:

The use cases are intended to be characteristic of classes of scenarios, and needn't be a specific real-world case, but including realistic elements can help make the use case concrete for readers, and tying to real world type cases they can relate to.

Future needs

At this time we have created a template for the use cases, we have a list of potential cases we would like to see created, and we are working on adding finished cases to the list. More finished products and additional topic suggestions are needed. In the future, we would like to see that each item on this list is represented twice – from the user perspective and the data provider.

Ultimately engagement by the community is needed to achieve these goals.



<http://cweb2.loc.gov/service/ppp/cph/3b50000/3b52000/3b52000/3b52086r.jpg>

5. PCCS Artifacts

Definition: List of PCCS categories and associated content.

Example:

- Data usage information
- Informal feedback from users (e.g., Amazon-style comments)
 - Publications about the data
 - Publications that use the data
- Data "peer review" information. This is ill defined, but could include
- Audit information about practices and processes to produce and maintain the data
 - Advice from scientific advisory groups, etc.
 - Etc.
- Authority or certification information
- Who is the authority (if there is one) that is asserting that the data meet certain quality criteria (e.g. Nat. Weather Service)
 - Criteria used in the certification