Towards a Standard for Provenance and Context for Preservation of Data for Earth System Science

Hampapuram K. Ramapriyan and John Moses

NASA Goddard Space Flight Center

Rama.Ramapriyan@nasa.gov

Long-term data sets with data from many missions are needed to study trends and validate model results that are typical in Earth System Science research. Data and derived products originate from multiple missions (spaceborne, airborne and/or in situ) and from multiple organizations. During the missions as well as well past their termination, it is essential to preserve the data and products to support future studies. Key aspects of preservation are: preserving bits and ensuring data are uncorrupted, preserving understandability with appropriate documentation, and preserving reproducibility of science with appropriate documentation and other artifacts. Computer technology provides adequate standards to ensure that, with proper engineering, bits are preserved as hardware evolves. However, to ensure understandability and reproducibility, it is essential to plan ahead to preserve all the relevant data and information. There are currently no standards to identify the content that needs to be preserved, leading to non-uniformity in content and users' not being sure of whether preserved content is comprehensive. Each project, program or agency can specify the items to be preserved as a part of its data management requirements. However, broader community consensus that cuts across organizational or national boundaries would be needed to ensure comprehensiveness, uniformity and long-term utility of archived data.

The Federation of Earth Science Information Partners (ESIP), a diverse network of scientists, data stewards and technology developers, has a forum for ESIP members to collaborate on data preservation issues. During early 2011, members discussed the importance of developing a Provenance and Context Content Standard (PCCS) and developed an initial list of content items. This list is based on the outcome of a NASA and NOAA meeting held in 1998 under the auspices of the USGCRP, documentation requirements from NOAA and our experience with some of the NASA Earth science missions. The items are categorized into the following 8 high level categories: Preflight/Pre-Operations, Products (Data), Product Documentation, Mission Calibration, Product Software, Algorithm Input, Validation, Software Tools.

For each content item, it is important to provide a definition, description, rationale (why content is needed), criteria (how good content should be), priority (high, medium, low; or critical, essential, desirable), Source (who should provide content item), project phase for capture, User community that would benefit from preserving the content item, whether and which existing standard (e.g., SensorML, ISO 19115) covers the representation of the content item, and any restrictions on being able to obtain, preserve and share the content item.

Developing a standard requires active involvement and review by many groups. Primary groups that must have inputs before establishing the standard for provenance and context content are: producers of data and derived products, intermediaries responsible for archiving and distribution, end users, agency program managers and policy makers.