Constructing a Data Management Plan Framework

Stanley Smith[1], Steven Tessler[2], Michael McHale[3]
U.S. Geological Survey
[1]Alaska Science Center, [2]New Jersey Water Science Center, [3]Climate Effects Network

The Alaska Science Center (ASC) and the Climate Effects Network (CEN) share the need to create a Data Management Plan (DMP) that provides guidance for researchers, facilitates data integration between multiple programs, and provides for the long-term preservation and availability of our valuable data assets. Over the past year the ASC and CEN researched current operations, participated in draft writings, and held numerous discussions to identify the 'best practices' in data management for research organizations. This proposed concept for developing a DMP framework (DMPf) is the result of that work. The intent of this document is to describe a framework for data management plans that will accomplish all the stated objectives and additionally be easily adaptable for other programs.

During the past 10 months the ASC-CEN data management team has researched and collected numerous data management documents in print and on the web from various federal agencies, academia, and other organizations both domestically and internationally. The team also made site visits to discuss data management best practices currently in operation at the National Environmental Observing Network (NEON), the National Park Service's Inventory and Monitoring Program (NPS I&M), the USGS Western Ecology Research Center (WERC) in San Diego, the USGS Fort Collins Science Center, the USGS Earth Resources Observation and Science (EROS) Center, and others. Although many data management practices are shared by all, the approach taken to document these practices has been unique within each organization. Each organization produced DMPs that supported their mission, but none are sufficiently generalized in context, coverage, or document structure to be re-used by other programs without significant modification.

From this beginning and informed by our new partnerships and literature review, we recognized the need for a generalized data management plan framework for research-based organizations. The challenge is to construct a basic framework for DMPs that could be customized for specific programs or projects, while at the same time accommodating the data management and data integration goals of the organization. This basic framework must incorporate the best practices used by the USGS and similar organizations, and be designed to be easily adaptable by multiple projects and programs within the USGS. The construction of the DMPf will make considerable use of the library of existing documentation we have collected.

For the above objectives to be realized, the DMPf must recognize three over-arching concepts - or 'dimensions'. These concepts are: Dual Data Management Cycles, Data Levels, and Document Layering.

Dual Data Management Cycles

The first framework concept is called "Dual Data Management Cycles". This concept grew from our direct observation within the USGS of the disparate data management needs between researchers and traditional data management professionals. Research organizations such as the USGS need to manage data for two separate purposes: to support the immediate and flexible needs of researchers, and to provide for the long-term preservation and wide accessibility of data. The 'dual cycle' concept acknowledges the legitimate needs of both perspectives and expresses this as two separate but tightly linked data cycles: the research data cycle and the long-term preservation data cycle. These two cycles serve different needs and require different data management skills and approaches.

Research Data Cycle

In the research data cycle researchers need substantial control over the data, flexibility for handling and reshaping data, and the ability to use familiar tools (such as spreadsheets) for manipulating and managing their data. They also need to ensure that the data fit the requirements of their chosen analytical tools, and that confidentiality is maintained pending publication.

Preservation Data Cycle

In contrast, the long-term preservation data cycle is more the concern of the parent organizations, investigators, and future researchers. The preservation cycle is generally overseen by data management professionals who build and maintain data systems, adhere to standards, maintain documentation, transition data with changing technologies, and whose goals are to keep the data in its original context while providing accessibility to both current and future users.

The contrasting missions between research and long-term data management can create tension in an organization when the needs of one cycle are pitted against the needs of the other – especially in a funding-limited environment. A DMPf must recognize and address the needs of both of these cycles, and by doing so create a near seamless transition from the short-term research data cycle to the long-term preservation and delivery of those data.

Data Levels

The second concept needed to build a robust data management framework is that of "Data Levels". A generalized DMPf must support data migrating from their original collection state through various 'Levels' of processing and organization. In this context data levels are used to identify the degree of transition from raw data to data products to published science documents. Each change in state brings with it a change in data management requirements.

This concept was borrowed from NEON (1) who in-turn borrowed it from NASA (2). These organizations designed their 'Levels' of transition to meet their unique needs, and the levels defined below are those we propose for USGS. These levels were defined to optimally correspond with changes in data management requirements as the data – and their progeny products - transition from initial measurement to research products. Without the benefit of this differentiation, discussions on data management, metadata, data formats, and data access often result in misunderstanding or disagreement because the topics themselves are level-specific. Recognizing and documenting data at different levels also provides the language necessary to trace higher level data products back to their source data. In the proposed DMPf we recognize data transition as comprising five Levels:

Level-0 - Raw Data

Level-0 or 'raw' data refers to data which are unaltered from their original collection, except that hand-written records have been transcribed or copied and stored in an electronic format. Level-0 data may exist as discrete values recorded by a field worker, data streamed from instruments, photographs or traces, or tissue biopsies. The critical indicator is that the data are in an unaltered condition as collected at the source. Raw data is not usually available for direct use. However, Level-0 data are a high priority for stable, long-term storage because data from all other levels are derived from Level-0 data.

Level-1 - Base Data

Level-1 or 'base' data refers to Level-0 data which have been reviewed and edited for quality. Base data are effectively data that have been certified as ready-to-use. Activities that transition Level-0 to Level-1 data include flagging values for quality issues; removing outliers; handling missing data as appropriate; standardizing field names, units and formats; rearranging data; aggregating data into collections with similar data; and making collections of data accessible. In some cases well-formed Level-0 data may simply become certified as Level-1 quality and undergo no change at all. The result of these activities produces the high-quality, accepted, base facts used in all higher level interpretations of the original data. Base data may exist in many forms, such as discrete facts in relational databases, processed scenes from satellites, or LIDAR delivered by a contractor.

Level-2 - Data Products

Level-2 data are 'data products' resulting from further transformation of Level-1 data, or combinations of Level-1 and Level-2 data. Level-2 data are produced when base data are transformed by some degree of processing into formats and derived products that are more convenient for specific groups of data consumers. Level-2 transformations can range from the simple computation of stream discharge from stage data, to the complex calculations needed to transform Level-1 or other Level-2 data into seamless images, grids, or NetCDF files.

Level-3 - Project Databases

Level-3 or 'project databases' are the databases built by researchers as they accumulate data from all other levels to conduct their research. The main criterion defining Level-3 data is that the data have been integrated into a combined dataset for project use. This includes spatial and temporal registration to standards, blending data from different sources internal and external, and other similar activities. Proper data management at lower levels can mitigate some of the requirements for Level-3 data integration, but will not eliminate them.

Level-4 - Project Products

Level-4 or 'project products' are any product, publication, or document produced from the Level-3 research database. Much of data management at this level recognizes the essential nature of 'records' or 'document' management in a comprehensive data management framework. All project products should demonstrate productive use of project funds by being cataloged and discoverable, and easily accessible to other research, funding agencies, cooperators, and the public through data repositories. Project products may also include project databases made available to support specific data products such as a journal articles, USGS reports, or National Science Foundation projects.

Inclusion of Levels in the DMPf is important because it requires us to consider data management throughout the full evolution of data from its initial collection to publication, and provides traceability from higher level data products back to their original source data. Data Levels also help to focus our

thinking as we address broad concepts such as metadata, operational procedures, and data structures because the requirements change from level to level.

Document Layering

The third dimension for the DMPf is "Document Layering". This concept was borrowed from the NPS Inventory and Monitoring program's guidelines for National Park DMPs (3). We define layering as the process of differentiating Agency principles, standards, and requirements from those unique to a specific program or project. Layering creates a framework that sets the foundation for data integration across an agency or program. Science Centers, programs, and projects will be able to use a common framework to produce a DMP tailored to their specific requirements with minimal effort, while acknowledging and adhering to the broader requirements of an organization or data-sharing program.

Four layers are currently defined in our DMP document stack. The first two layers are common to all DMPs and the next two are specific to individual DMPs.

Base Layer
The first or 'base' DMP Layer provides the common DMP structure pre-populated with a Table-of-Contents and all of the common sections completed. The text of this layer provides the purpose, theory, and objectives of the DMP including the definition of roles and responsibilities, common vocabulary, case studies, references, etc.

Standards and Integration Layer
The second DMP Layer includes standards and procedures that will facilitate the integration and exchange of data across all DMPs written using this framework. These include sections for data standards, references to collection protocols, metadata standards, data model standards, naming conventions, data exchange formats, web service guidelines and other processes or requirements necessary for data integration.

Local Implementation Layer
The third DMP Layer documents requirements unique to a science center, program, or project. This is the only Layer an entity using the framework would need to complete in order to create their program-specific DMP because the 'base' and 'standards' layers are already complete and define the basic requirements shared by all data partners. The Local Implementation Layer documents any extensions to the DMP that the entity writing their program-DMP would require above and beyond base requirements. Deviations from the framework standards are also documented and defended in this layer.

Tactical Planning Layer
This last document Layer contains the year-by-year tactical plans for a data management program. This Layer is optional based upon the needs of the program to have an implementation schedule outlined within the DMP.

Summary

In summary, the 'Dual Management Cycle' acknowledges and clarifies the different roles researchers and data management professionals play in the development and preservation of data resources, 'Data Levels' defines a context for discussing handling and processing of data and ensures that data are

managed and traceable throughout their entire life span, while DMP 'Document Layering' creates a uniform template for reusability across many separate implementations, while facilitating data integration.

Development of the DMPf

An economy and broader adoption can be achieved by collaborating in the development of the DMPf with data management professionals across the USGS. The ASC and CEN have committed to jointly developing the DMPf just described. Our desire, however, is to expand the team to include contributors from other USGS programs and potentially other DOI Bureaus with similar needs. A broader team would help to create a truly generalized and portable DMP framework.

Much of the material for this proposed framework already exists in data management plans developed by research organizations and federal agencies from the U.S. and abroad. Wording from these existing plans can be adopted, adapted, and used to produce a general framework that will serve much of the USGS and other federal partners. After all, data management in support of research is not something new.

Works Cited

1. **NEON.** *NEON Cyberinfrastructure Architecture.* Cyberinfrastructure Team, National Ecological Observatory Network. 2010. Version B Draft (Not Published). NEON.DSDV.CYI.003024.DSN.

2. **NASA.** Earth Science Data Terminology and Formats - Data Processing Levels for EOSDIS Data Products. *NASA Science - Earth.* [Online] http://science.nasa.gov/earth-science/earth-science-data/earth-science-data-terminology-and-formats/.

3. **NPS.** *Data Management Guidelines for Inventory and Monitoring Networks.* Natural Resource Program Center, National Park Service - Department of the Interior. 2008. Natural Resource Report NPS/NRPC/NRR-2008/035. [Online] http://science.nature.nps.gov/im/datamgmt/docs/DMPlans/National_DM_Plan_v1.2.pdf.