A PROPOSAL SUBMITTED TO:
USGS Community for Data Integration

USGS Science Center Adaptable Data Management Plan Implementation

By Thomas E. Burley[1] and Stan W. Smith[2]

[1] USGS Texas Water Science Center
[2] USGS Alaska Science Center

Summary

Data management and data integration have been identified in the USGS science strategy as critical areas that are essential to the success of future USGS science (U.S. Geological Survey, 2007). USGS science centers falling under all mission areas have largely operated under their own purview in the arena of data management with the level of oversight and consistency among projects varying greatly. Water science centers manage a considerable number of local, regional, and national projects in cooperation with Federal and State partners that produce data that often fall outside the interests of national USGS data Programs such as the North American Water Quality Assessment (NAWQA) and the National Water Information System (NWIS). In addition, a significant increase in new data types and formats (for example, geospatial data, lightweight databases, and integrated data originating from multiple sources) have created a need for more formalized planning to ensure that data-driven projects are efficient and that the data being produced across science center projects are consistent and well documented. To this end, development of an adaptable data management template to construct science center data management plans from a common framework will help Program-level offices improve their data business practices while also integrating Program planning with enterprise-level objectives.

Background

The goal of the USGS Science Strategy is integration of the vast capabilities of the Bureau to better serve a Nation facing significant natural science and societal challenges (U.S. Geological Survey, 2007; Burkett and others, 2011). Multiple approaches exist to achieve this end. But underlying all Bureau integration strategies is the need for accessible and high confidence data and information from the science community. To address this issue and meet the needs for science, the Bureau Science Strategy Team highlighted the accessibility of data across multiple disciplines, geographic, temporal, and political boundaries as a fundamental enabling objective.

To better understand the breadth of the data management requirement of the Bureau, consider that virtually all research and analysis involve some data and that the Department of the Interior's Federal Information Systems Security Awareness program (FISSA) estimates that "more than 60% of an organization's economic assets are information assets in the 21st century" (Department of the Interior, 2011). Practical data management facilitates understanding and usability, preserves provenance and context, and results in data products that are highly reliable, well-organized, well-documented, accessible, and user-friendly. Well-managed data become long-term asset for the Bureau and for the public. A professional and consistent approach to data management ensures that the maximum value is derived from each sampling or research effort, and helps to justify the cost of those efforts with the ready availability of high-quality data products.

The negative costs and reduced utility associated with lack of data and information management are very real. The University of Tennessee Institute for a Secure and Sustainable Environment conducted a two year study on the impacts data management practices have with multi-agency natural resource management efforts. The state of priority legacy datasets associated with the management of a high-elevation ecosystem as well as the business costs associated with past and present activities in the context of data governance were examined. The study estimated a cost of over $112,000 associated with inadequate data and information management which did not

include significant in-kind contributions as well as initial study implementation costs (Peine and others, 2008).

Data management at the USGS Program level has varied greatly in terms of extent and consistency of data management planning, acquisition, processes, standards, accessibility, and documentation. As an example, a considerable number of water science center cooperator projects produce important data that do not fall under the umbrella of the NAWQA or NWIS Programs. This has resulted in numerous fragmented and orphaned datasets. In addition, the lack of Enterprise-level guidance has left Program-level projects on their own and subsequently Program scientists feeling burdened with the magnitude of taking on data management at all levels. However, a paradigm shift is emerging with the initiation of the Core Science Systems science strategy which recognizes the importance of data and information management and serves to elevate and expand the investment in this area.

A consistent and unified approach to data management across the Bureau is critical for ensuring that progress we make today does not inadvertently recreate the same issues we face from the past. In addition, the USGS will be better served to draw upon successful existing data management programs both in the Federal and International communities. The USGS Climate Effects Network (CEN) and the Alaska Science Center (ASC) have initiated a significant effort doing just this to lay out a common *framework*, the Data Management Plan Framework or DMPf, by drawing on the greater data management community as well as determining what was currently being done within the Bureau (Smith and others, 2011). By devising a common framework from which Program-level and Project-level plans can be developed that also incorporates the larger enterprise goals and policy, the burden of data management and data integration on science centers and scientists is lessened and data integration is promoted.

A document layering approach for the DMPf was adopted from the National Park Service (NPS) Inventory and Monitoring Program's guidelines for National Park Data Management Plans (National Park Service, 2008). The advantage of a layered approach is that it separates specific Project and Program-level material from guidelines that apply across the Bureau. Duplication of effort is reduced, consistency is enabled, and data integration is promoted in that lower level documents inherit the higher level requirements, vocabulary, and standards. The DMPf is a living document and Bureau data management teams are invited to enhance and extend this framework. An outline of the principle layers in version 1 are shown below in fig. 1.

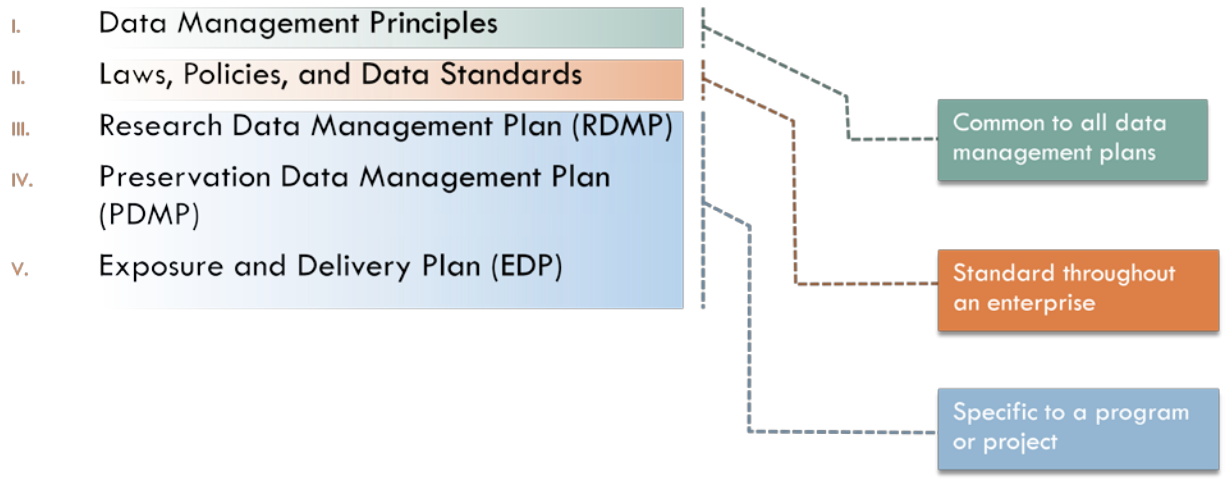

Figure 1: Data management plan implemented as document layers (adapted from the National Park Service Inventory and Monitoring Program) including separation of Research and Preservation planning (as recommended by UK Data Archive).

The DMPf further incorporates the concept of "data levels" to describe the states of maturity as data is successively transformed during a project lifecycle. Recognition of data level ensures that the various states (raw, base, derived, final product) are each managed according to appropriate practices. Data level differentiation is important to the understanding of data management in that the needs for documentation, metadata, data formats, and data access vary with each data level (fig. 2). The concept of data levels was borrowed from NEON (National Ecological Observation Network, 2009).

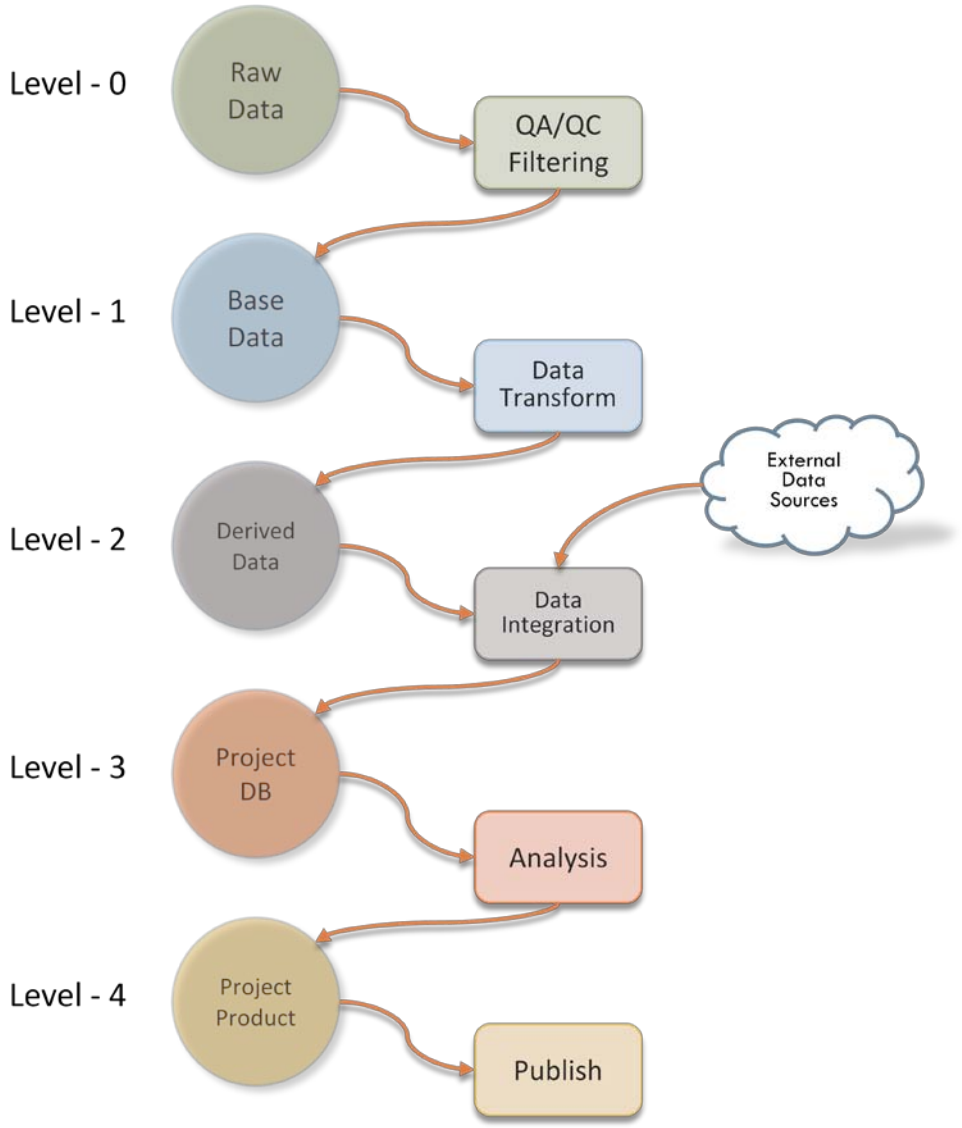

Figure 2: A simplified conceptual representation of the various levels of data management (adapted from the National Ecological Observation Network (NEON) who in turn borrowed it from the National Aeronautics and Space Administration (NASA)).

The DMPf is a living document and serves as a head-start for developing Program and Project-level data management plans and its revision and expansion will be further informed by the development of science center level data management plan implementations.

Problem

USGS integrated and water science center Programs along with a myriad of new data types and technical considerations requires a more formalized and consistent approach to Program and Project-level data management.

Objectives and Scope

The study has three primary objectives:

1. Understand the test case science center business models, roles and responsibilities, requirements, and data governance needs.
2. Develop Program level templates for two USGS business models (water science centers and integrated science centers) based on the DMPf version 1 framework. The water science center template will be modeled on the Texas Water Science Center (TXWSC) and the integrated science template will be modeled on the Alaska Science Center (ASC). Both templates are intended to be reusable and extensible to other USGS science centers using the water or integrated science center business models.
3. Make available internally to USGS a wiki version of the DMPf to enhance future participation and development.

Approach

Task 1– Program and Project Business Process Assessment

Working with selected TXWSC staff and management, the objectives of this task will be:

(a) Examine the Texas Water Science Center and Alaska Science Center business models, projects, and data requirement characteristics to gain an understanding of the similarities and differences between water and integrated science center data management requirements.

(b) Identify data management roles and responsibilities. Employ the principle of inheritance to separate roles and responsibilities appropriate to science center Programs and those appropriate to their Projects.

(c) Identify historical science center data issues and roadblocks by soliciting input from senior science center staff in order to inform both the DMPf and the science center plan outline.

(d) Develop and prioritize Program data governance goals and performance measures drawing upon Task 1(a) through 1(c) information.

(e) Conduct literature review of additional existing Department of Interior agency and external data management plans.

INTERIM DELIVERABLE: Summary of key issues persisting in science center Program, roles and responsibilities necessary to success, historical roadblocks as hindsight, and Program goals and measures which will be used to revise or expand the existing DMPf draft.

Task 2–Development of an Annotated Science Center Data Management Plan Outline

This task will draw upon Task 1 as well as Layer-III (Research Data Management Plan - RDMP) of the DMPf to develop an annotated science center RDMP development guide.

The RDMP development guide will consist of two separate RDMP development templates to accommodate the business model differences between integrated science centers and water science centers. Key elements including roles and responsibilities, technology, standards, and services critical for Program and project-level success will be incorporated. Key staff will be queried to identify existing resources to ensure this effort can build upon what already exists. In spite of differences, it is expected that the two business models will have significant overlap. In keeping with the inheritance concept designed into the DMPf, overlapping practices will be lifted into a new Bureau level (USGS) layer that both the water and integrated science center templates inherit.

The development guide of this task will serve as the reusable science center RDMP development template. This information will also be used to re-inform and further develop the working DMPf draft.

DELIVERABLE: Two RDMP development guides (Integrated Science Center and Water Science Center) and a parent USGS RDMP development guide. These deliverables will be subsequently used to draft the two specific science center RDMPs in Task 3.

Task 3–Development of a Version 1 Science Center Plan Implementation

Using the annotated templates developed during Task 2, a full version 1 implementation of the RDMPs will be developed for the Texas Water Science Center and Alaska Science Center. Key science center staff and management will be asked to review and provide input which will be used to revise the plan and the Task 2 reusable plan template. The development of two separate plans will be necessary given the inherent business model differences between integrated and water science centers and the level of detail needed to provide adequate direction to Programs and their Projects.

Upon completion of the version 1 implementation, feedback will be gathered as result of a review with selected management and staff from the TXWSC and the ASC to ensure the implementation reflects key science and business model considerations and needs.

DELIVERABLE: Two science center plan implementations based on the DMPf.

Task 4–Development of a Wiki version of each Science Center Data Management Plan Implementation

For a data management plan to be successful it must be a living document. That is, frequent review and updating of the plan must occur to ensure that it accurately represents Program and Project-level business requirements. A wiki implementation of the data management plans will promote Program data manager and scientist involvement in the revision of the plan. In addition, the interlinked page design of a wiki implementation of the plan allows for more natural and meaningful topic associations among plan wiki pages.

DELIVERABLE: A wiki implementation of each science center data management plan in an appropriate wiki software package (for example, MediaWiki or other appropriate wiki platform).

Quality Assurance Plan

Input will be gathered from science center staff and management through iterative reviews and input at each task to ensure that the data management plan takes into account science center business processes and that goals and measures, roles and responsibilities, and standards are realistically defined.

Relevance and Benefits

This study will advance the long-standing need for a more formalized approach to data management planning at the science center level in USGS by one more step. The study will use two different science centers as test cases. Improved planning for data management and data integration is identified in the Bureau science strategy goals referenced herein with the need for consistent and unified data management to allow for accessible and high confidence data and information from the USGS science community.

Deliverables

The primary deliverables will include two science center plan implementations based on the DMPf available as reusable templates for other USGS science centers along with wiki implementations of the DMPf for use internally within the CDI community.

Timeline

Tasks	Number of months after receipt of funding						
	1	2	3	4	5	6	7
Task 1 - Program and Project Business Process Assessment							
Task 2 - Annotated Science Center DMP Outline							
Task 3 - Develop V.1 Science Center Plan Implementation							
Task 4 - Develop Wiki Version of Science Center DMP							

Budget

The total amount requested is $68,500.

References

Burkett, V.R. and others, 2011, Public review draft; USGS global change science strategy: A framework for understanding and responding to climate and land-use change, U.S.

Geological Survey Open-File Report 2011–1033, 32 p., at http://pubs.usgs.gov/of/2011/1033.

Department of the Interior, 2011, Federal Information Systems Security Awareness + Privacy and Records Management (FISSA+), Course Code: DOI-ITSec-Ann-LCMS.

National Ecological Observation Network, 2009, The NEON data product concept and production plan, 2 p. Available at: http://www.neoninc.org/sites/default/files/NEON_Data_Product_Concept_and_Production_Plan.Mar2009_0.pdf

National Park Service, 2008, Data management guidelines for inventory and monitoring networks: Natural Resource Report NPS/NRPC/NRR—2008/035, National Park Service, Fort Collins, Colorado, 126 p. Available at: http://science.nature.nps.gov/im/datamgmt/docs/DMPlans/National_DM_Plan_v1.2.pdf

Peine, J.D., T.E. Burley, B.L. League, S.C. Hetrick, 2008, Data Management: A Critical Link Between Scientists and Managers, Proceedings of the Society for Conservation Biology Annual Meeting, Chattanooga, TN. Available at: http://www.conbio.org/activities/meetings/2008/program/MONDAY_14_July.pdf

Smith, S.W., S. Tessler, M. McHale, 2011, United States Geological Survey Data Management Plan Framework (unpublished).

U.S. Geological Survey, 2007, Facing tomorrow's challenges—U.S. Geological Survey science in the decade 2007–2017: U.S. Geological Survey Circular 1309, x + 70 p. Available at: http://pubs.usgs.gov/circ/2007/1309/