Science Center Data Management Plan Framework (DMPf): Perspectives, Findings, and Progress

Stan Smith – Alaska Science Center (ASC)

Tom Burley – Texas Water Science Center (TXWSC)

Steve Tessler – New Jersey Water Science Center & National Water Census (NWC)

U.S. Department of the Interior
U.S. Geological Survey

CDI Webinar Sept. 5, 2012

Relevance to CDI and Science Support Framework

- Supports USGS Data and Information Management
- Contributes to interdisciplinary science and promotes standards and best practices
- Crosswalks directly to the Science Data Life Cycle

Background and Problem

- USGS has more projects than ever, all generating or using data
- Data exist in an increasingly complex array of types and formats
- Workflows involving data handling vary widely among projects and staff
- “Data management” (DM) is assumed and rarely itemized as a part of a project’s management, and no specific DM guidance exists

- We have a Corporate need to locate and access our wide range of data assets for both emergency responses and other activities in support of science, and to preserve data for future use regardless of whether or not it has a formal ‘home’ (NWIS, etc.)

➢ To arrive at an integrated data future, we need a more formalized and consistent approach to Program- and Project-level data management.

CDI Project Deliverables & Benefits

- Summary of TXWSC (Center-based) issues, perceptions, and forward-looking goals
  - Communication; Understanding; Needs
- DMPf Research Data Management Plan (RDMP) - Enterprise Template, compatible with the needs of the TXWSC, ASC, and NWC
  - Respects phases of research
  - Detailed; Flexible for “À la carte” use by others
- Version 1.0 *beta* Program-level RDMP documents for the TXWSC and ASC based on the Enterprise document

Program Perspectives

- **Texas Water Science Center (TXWSC)**
  - Most staff either Data Program or Studies
  - ~160 staff total; ~50 Studies staff across 4 state offices (Ft. Worth, Austin, San Antonio, Houston)
- **Alaska Science Center (ASC) and USGS Climate Science Centers (CSC)**
  - Integrated Science Center: Biology, Geography, Geology, Water
  - ~200 staff with offices in Anchorage, Juneau, Fairbanks
- **National Water Census (NWC)**
  - Multidisciplinary and national in scope, 6 major data subject areas
  - Need coordination of all activities to produce an integrated product

TXWSC: Staff & Management Perspectives

- **Implementation critical**
  - Elements: Program policy and guidance, roles, oversight and review mechanisms
  - Communication, education, consider culture
- **Recognized need for defined guidance among nearly all interviewed**
  - Planning, file structures, data flows, documentation, archiving
  - Consistency, quality, efficiency
- **Personal data management woes common**
- **Science Center management buy-in**

DMPf Origins and Collaboration

- **2010-11 – initial funding by USGS Climate Effects Network**
  - Mike McHale – Climate Effects Network and New York Water Science Center
  - Steve Tessler – New Jersey Water Science Center
  - Stan Smith – Alaska Science Center and Changing Arctic Ecosystems Initiative
- **2012 – CDI funding to develop the DMPf for Center-based (Program) DMPs - to Tom Burley and Stan Smith (Steve Tessler participating 'in kind')**
  - Tom Burley – Texas Water Science Center, joined the development team
- **Objectives:**
  - Develop DMP template so it can be used as the basis for multiple programs and projects
  - Lay foundation for best practices, standards, and GS-wide data integration
- **Influenced by:**
  - Existing DMPs and related documents from government agencies and research organizations world-wide.
  - Boots-on-the-ground Science Center and Program perspectives

DMPf – 3 Core Concepts

1. Split Research from Preservation

- *Proposed by UK Data Archive, and others*
- Message to Researchers - “You don’t have to do it all”

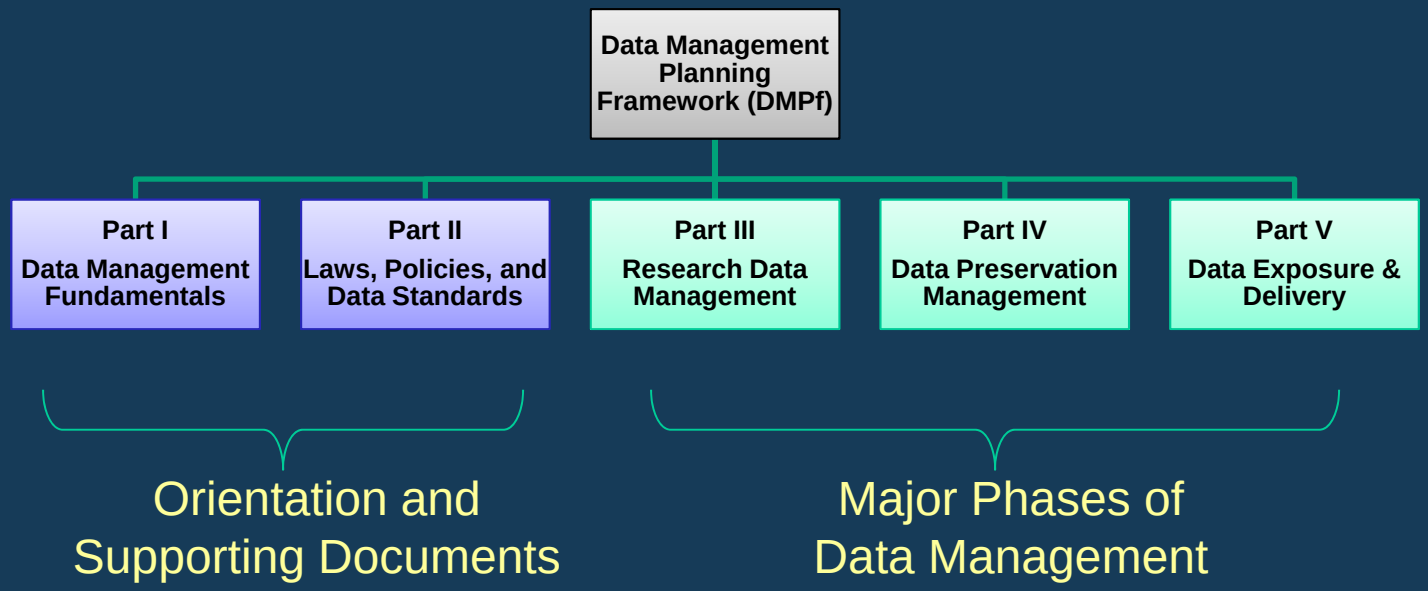
2. Recognize Data Levels

- *Used by NASA, NEON, and others*
- Raw Data, Base Data, Derived Data, Project Database, Data Products

3. Document Layering

- *Used by NPS Inventory & Monitoring Program*
- Overall DM Inheritance, Enterprise requirements, Program requirements, Project requirements

DMPf Structure – 5 Parts



The top of the DMP framework lays out the components of DM planning and policy.

Data Management Fundamentals need only be defined and rationalized once. All work adheres to the same laws, policies, and standards at the highest level.

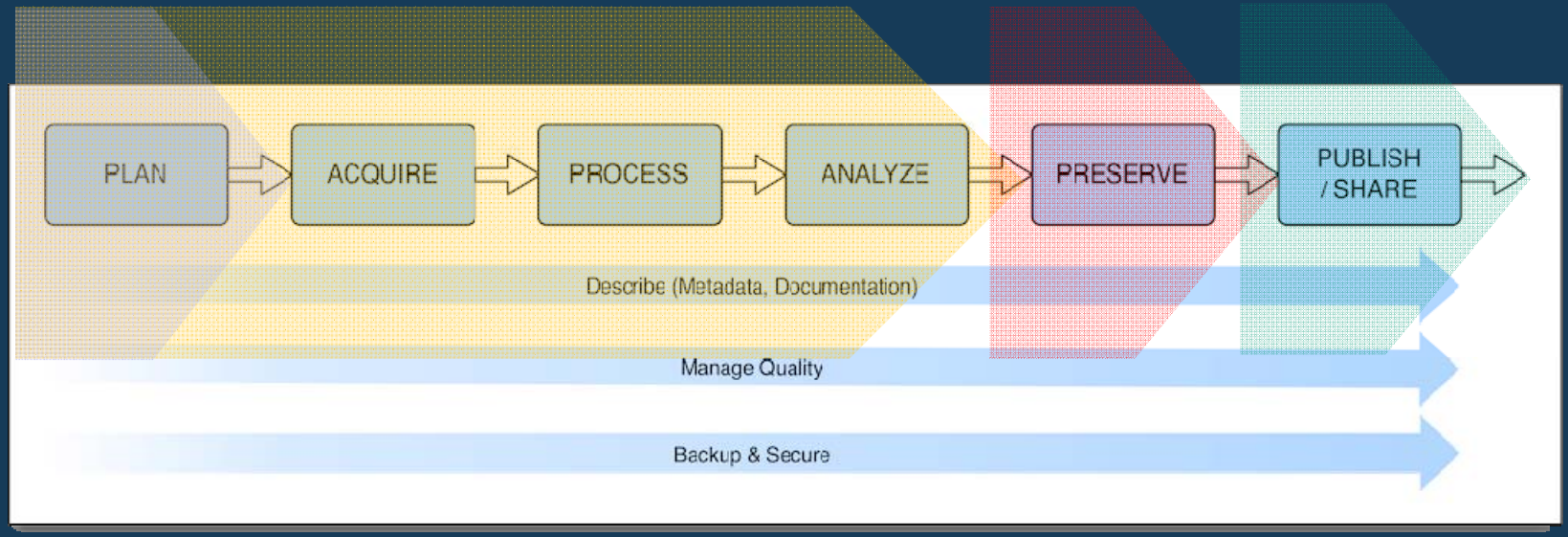
The Research process is declared distinct from Preservation and the Sharing of data.

DMPf and the Science Data Life Cycle

Part III
RDMP

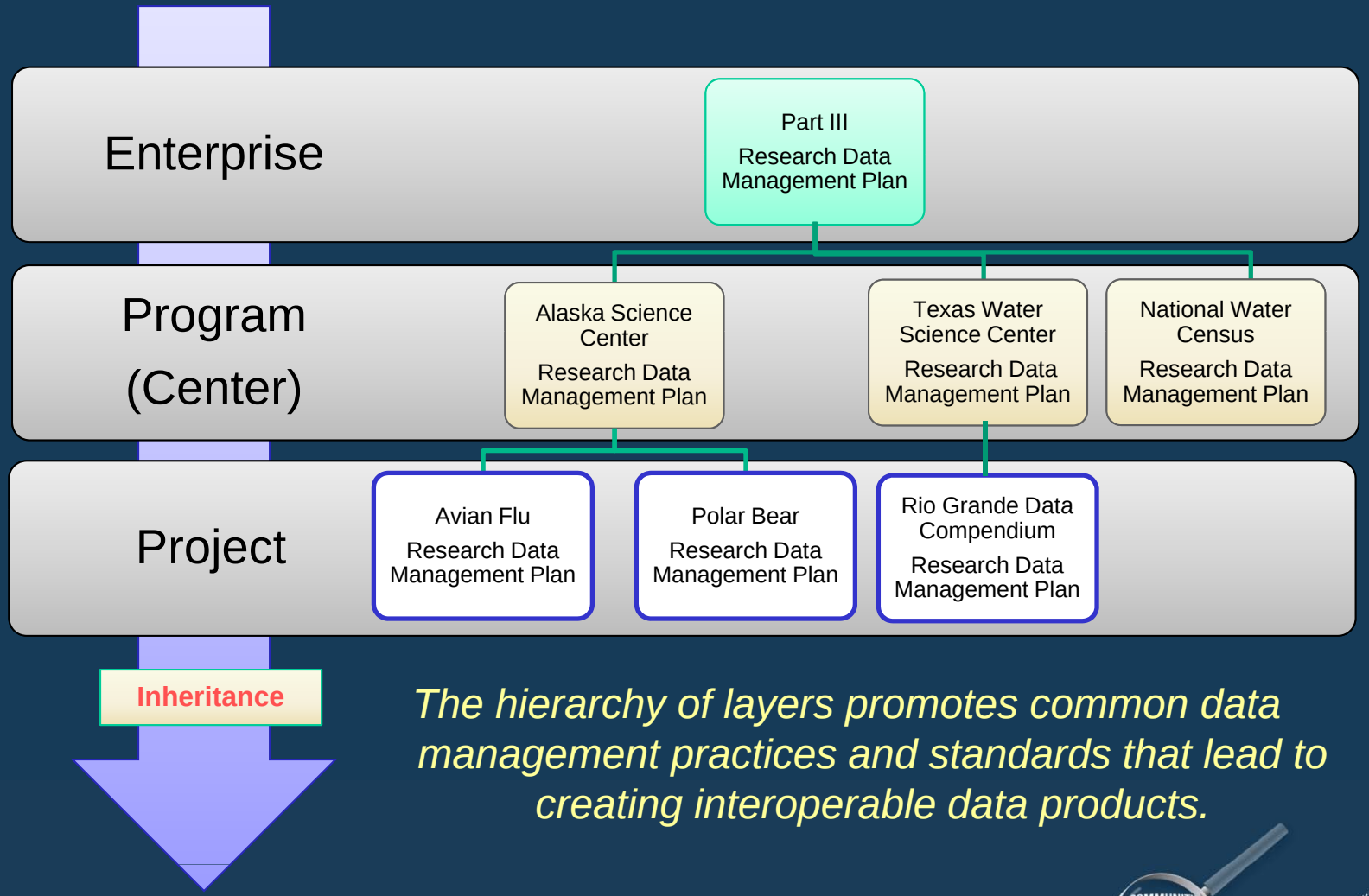
Part IV
DPMP

Part V
DEDP



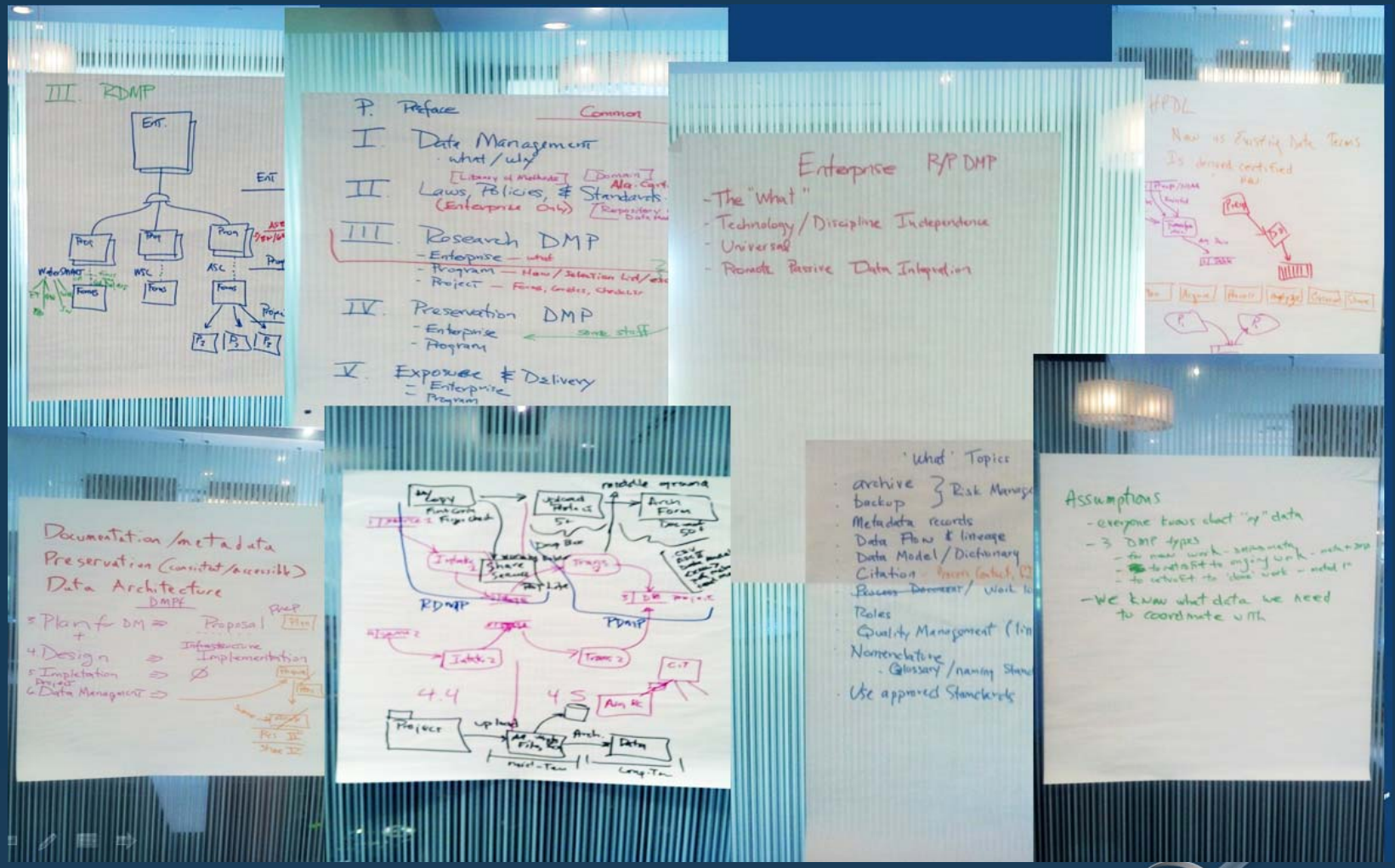
RDMP = Research Data Management Plan
DPMP = Data Preservation Management Plan
DEDP = Data Exposure and Delivery Plan

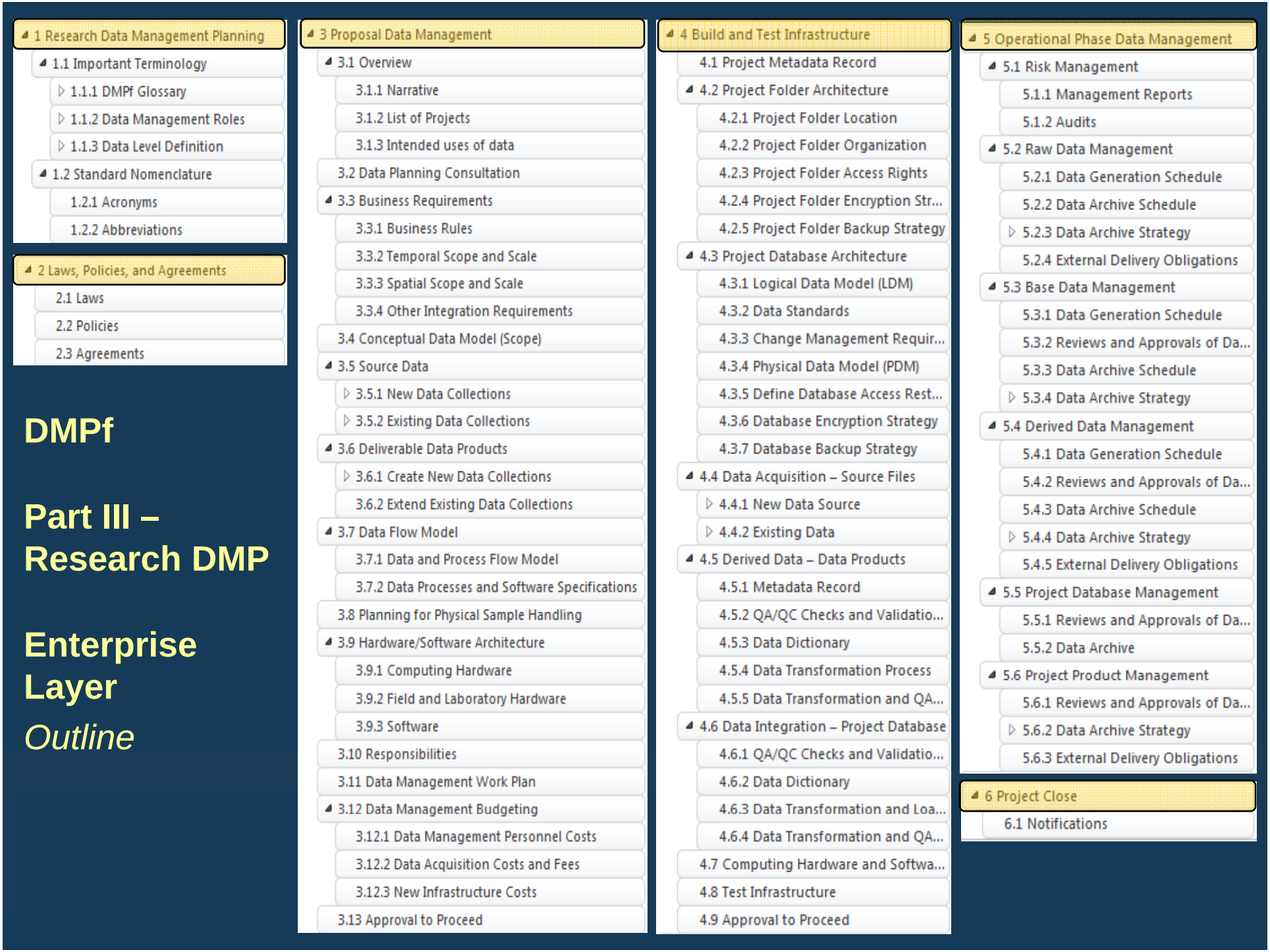
DMPf Part III – Document Layering



The hierarchy of layers promotes common data management practices and standards that lead to creating interoperable data products.

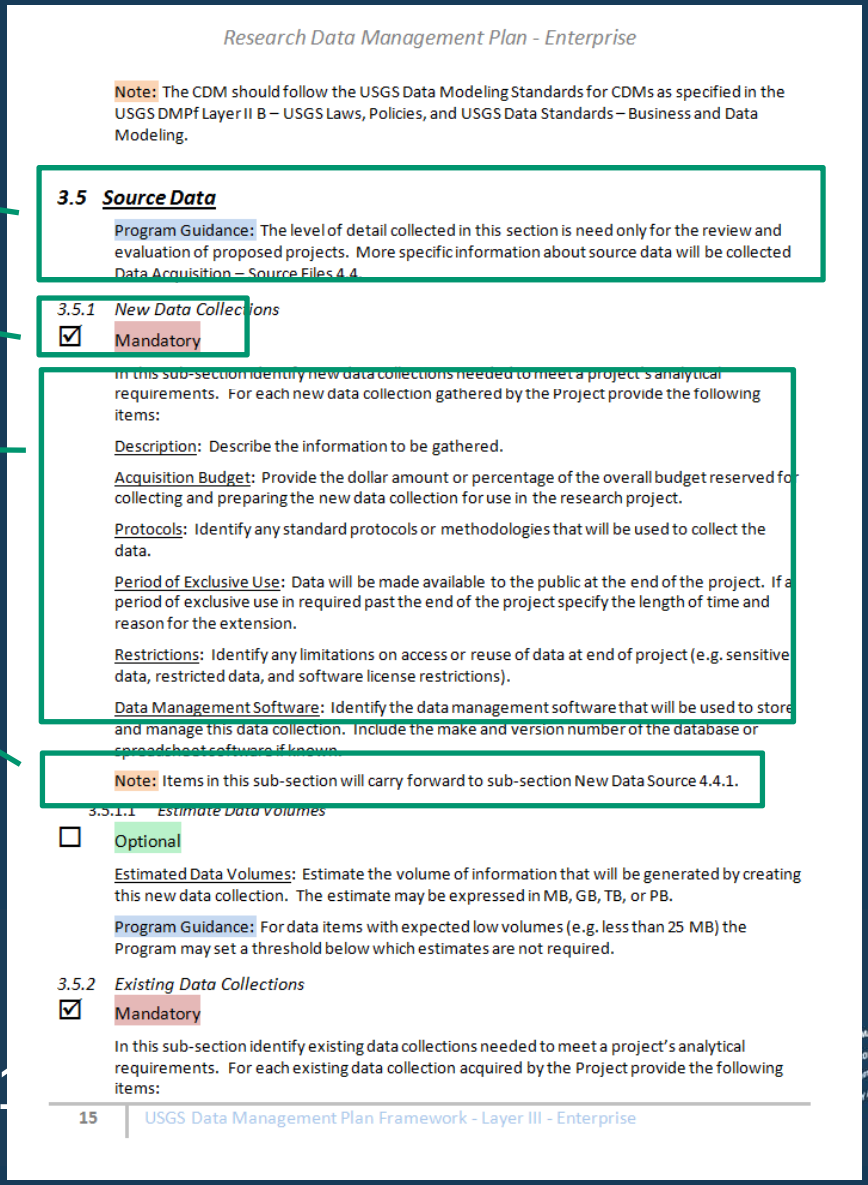
DMPf Boiler Room: July 9-13, 2012




1 Research Data Management Planning
1.1 Important Terminology
1.1.1 DMPf Glossary
1.1.2 Data Management Roles
1.1.3 Data Level Definition
1.2 Standard Nomenclature
1.2.1 Acronyms
1.2.2 Abbreviations
2 Laws, Policies, and Agreements
2.1 Laws
2.2 Policies
2.3 Agreements
DMPf
Part III – Research DMP
Enterprise Layer
Outline
3 Proposal Data Management
3.1 Overview
3.1.1 Narrative
3.1.2 List of Projects
3.1.3 Intended uses of data
3.2 Data Planning Consultation
3.3 Business Requirements
3.3.1 Business Rules
3.3.2 Temporal Scope and Scale
3.3.3 Spatial Scope and Scale
3.3.4 Other Integration Requirements
3.4 Conceptual Data Model (Scope)
3.5 Source Data
3.5.1 New Data Collections
3.5.2 Existing Data Collections
3.6 Deliverable Data Products
3.6.1 Create New Data Collections
3.6.2 Extend Existing Data Collections
3.7 Data Flow Model
3.7.1 Data and Process Flow Model
3.7.2 Data Processes and Software Specifications
3.8 Planning for Physical Sample Handling
3.9 Hardware/Software Architecture
3.9.1 Computing Hardware
3.9.2 Field and Laboratory Hardware
3.9.3 Software
3.10 Responsibilities
3.11 Data Management Work Plan
3.12 Data Management Budgeting
3.12.1 Data Management Personnel Costs
3.12.2 Data Acquisition Costs and Fees
3.12.3 New Infrastructure Costs
3.13 Approval to Proceed
4 Build and Test Infrastructure
4.1 Project Metadata Record
4.2 Project Folder Architecture
4.2.1 Project Folder Location
4.2.2 Project Folder Organization
4.2.3 Project Folder Access Rights
4.2.4 Project Folder Encryption Str...
4.2.5 Project Folder Backup Strategy
4.3 Project Database Architecture
4.3.1 Logical Data Model (LDM)
4.3.2 Data Standards
4.3.3 Change Management Requir...
4.3.4 Physical Data Model (PDM)
4.3.5 Define Database Access Rest...
4.3.6 Database Encryption Strategy
4.3.7 Database Backup Strategy
4.4 Data Acquisition – Source Files
4.4.1 New Data Source
4.4.2 Existing Data
4.5 Derived Data – Data Products
4.5.1 Metadata Record
4.5.2 QA/QC Checks and Validatio...
4.5.3 Data Dictionary
4.5.4 Data Transformation Process
4.5.5 Data Transformation and QA...
4.6 Data Integration – Project Database
4.6.1 QA/QC Checks and Validatio...
4.6.2 Data Dictionary
4.6.3 Data Transformation and Loa...
4.6.4 Data Transformation and QA...
4.7 Computing Hardware and Softwa...
4.8 Test Infrastructure
4.9 Approval to Proceed
5 Operational Phase Data Management
5.1 Risk Management
5.1.1 Management Reports
5.1.2 Audits
5.2 Raw Data Management
5.2.1 Data Generation Schedule
5.2.2 Data Archive Schedule
5.2.3 Data Archive Strategy
5.2.4 External Delivery Obligations
5.3 Base Data Management
5.3.1 Data Generation Schedule
5.3.2 Reviews and Approvals of Da...
5.3.3 Data Archive Schedule
5.3.4 Data Archive Strategy
5.4 Derived Data Management
5.4.1 Data Generation Schedule
5.4.2 Reviews and Approvals of Da...
5.4.3 Data Archive Schedule
5.4.4 Data Archive Strategy
5.4.5 External Delivery Obligations
5.5 Project Database Management
5.5.1 Reviews and Approvals of Da...
5.5.2 Data Archive
5.6 Project Product Management
5.6.1 Reviews and Approvals of Da...
5.6.2 Data Archive Strategy
5.6.3 External Delivery Obligations
6 Project Close
6.1 Notifications

DMPf Part III - RDMP Enterprise

- **Guidance** to plan writers
- **Mandatory and Optional** items
- **Specifies minimum data** to be collected
- **Notes** providing additional information



Research Data Management Plan - Enterprise

Note: The CDM should follow the USGS Data Modeling Standards for CDMs as specified in the USGS DMPf Layer II B – USGS Laws, Policies, and USGS Data Standards – Business and Data Modeling.

3.5 Source Data

Program Guidance: The level of detail collected in this section is need only for the review and evaluation of proposed projects. More specific information about source data will be collected Data Acquisition – Source Files 4.4

3.5.1 New Data Collections

☑ Mandatory

In this sub-section identify new data collections needed to meet a project's analytical requirements. For each new data collection gathered by the Project provide the following items:

Description: Describe the information to be gathered.

Acquisition Budget: Provide the dollar amount or percentage of the overall budget reserved for collecting and preparing the new data collection for use in the research project.

Protocols: Identify any standard protocols or methodologies that will be used to collect the data.

Period of Exclusive Use: Data will be made available to the public at the end of the project. If a period of exclusive use in required past the end of the project specify the length of time and reason for the extension.

Restrictions: Identify any limitations on access or reuse of data at end of project (e.g. sensitive data, restricted data, and software license restrictions).

Data Management Software: Identify the data management software that will be used to store and manage this data collection. Include the make and version number of the database or spreadsheet software if known.

Note: Items in this sub-section will carry forward to sub-section New Data Source 4.4.1.

3.5.1.1 Estimate Data Volumes

☐ Optional

Estimated Data Volumes: Estimate the volume of information that will be generated by creating this new data collection. The estimate may be expressed in MB, GB, TB, or PB.

Program Guidance: For data items with expected low volumes (e.g. less than 25 MB) the Program may set a threshold below which estimates are not required.

3.5.2 Existing Data Collections

☑ Mandatory

In this sub-section identify existing data collections needed to meet a project's analytical requirements. For each existing data collection acquired by the Project provide the following items:

15 | USGS Data Management Plan Framework - Layer III - Enterprise

DMPf Part III - RDMP Project Template

Project-Layer RDMP templates can be constructed from the requirements established in Program RDMP Guidance

Data Inputs – Existing Data Collections

1	[Name of Collection]
Description:	Describe the existing data, model, etc. that will be used. If not known, please provide as much information as possible (e.g., remote sensing, global climate models, etc.).
Restrictions:	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.).

Data Inputs – New Data Collection

1	[Provide a brief name to describe new data collection]
Description:	Describe the information to be gathered and the scale (e.g., national, regional, landscape, etc.) of the data.
Data Management Budget:	Describe the proposal budget portion allocated for data management activities for the new data collected.
Protocols:	Identify any standard protocols or methodologies that will be used to collect the data, if available.
Quality Checks:	Specify the procedures for ensuring data quality.
Exclusive Use:	Project data and associated products should be available publically at the end of the project. If a request to limit access for a period of time after project completion is needed, identify the length of time and the reason for the extension. (Request cannot be more than two years.)
Restrictions:	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.).
Contact:	Provide a point(s) of contact if questions arise related to the data and associated products (name, email, and phone number).

Example borrowed from NCCWSC CSC RDMP

DMPf Part III - RDMP Status

Enterprise

Part III
Enterprise RDMP
Version 2.0 in final review process

Still to do:
- Develop starting point Project template library for Program RDMP authors

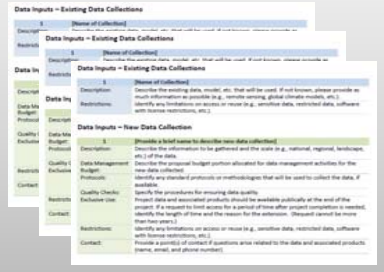
Program (Center)

ASC RDMP
ASC Management Team set to initiate September 17

TXWSC RDMP
Scheduled to initiate in September

NWC RDMP
Initiated

Project



Schedule for implementation of Project plans - TBD

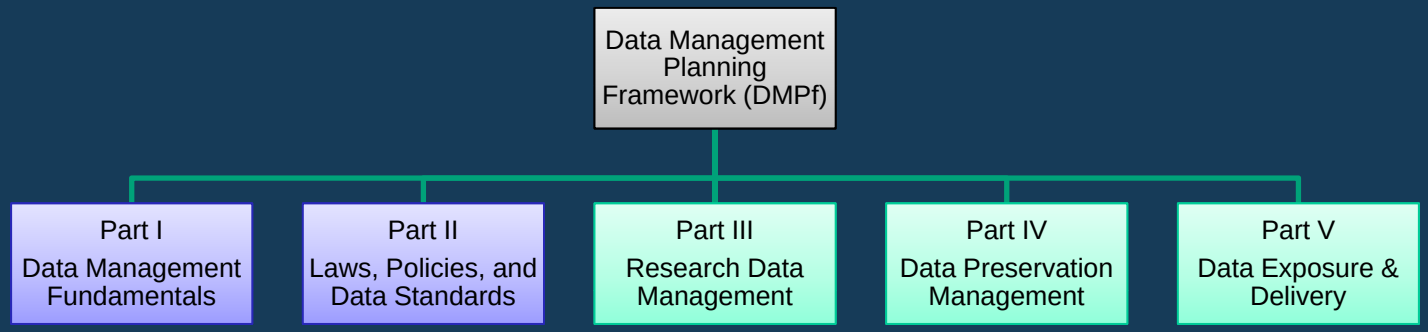
Science Center Data Management Plan Framework (DMPf): Perspectives, Findings, and Progress

Questions?

Contact:

- Stan Smith – stansmith@usgs.gov
- Tom Burley – teburley@usgs.gov
- Steve Tessler – stessler@usgs.gov

DMPf Next Steps



Ready to insert:

- Guide to ISO 19115-2
- Standard for temporal datatype (ISO 19108)
- Standard for spatial datatype (ISO 19107, 19111)

- Draft outline for Part IV - Enterprise
  - Archive
  - Backup
  - Disaster Recovery
  - Off-site storage
  - Data (file) up-loaders
  - Data integration
  - Center databases
  - Enterprise databases
  - DBA responsibilities