Data Management Planning

– from theory to practice in research data management

CDI/PNAMP Data Management Webinar, March 18, 2015

Stan Smith, USGS Alaska Science Center
Steve Tessler, USGS New Jersey Water Science Center

Two useful things to note

DMP - Data Management "Plan"

1. *Plan* is a misleading word

Proposal Stage: DMP *is* a plan, and a *promise*.

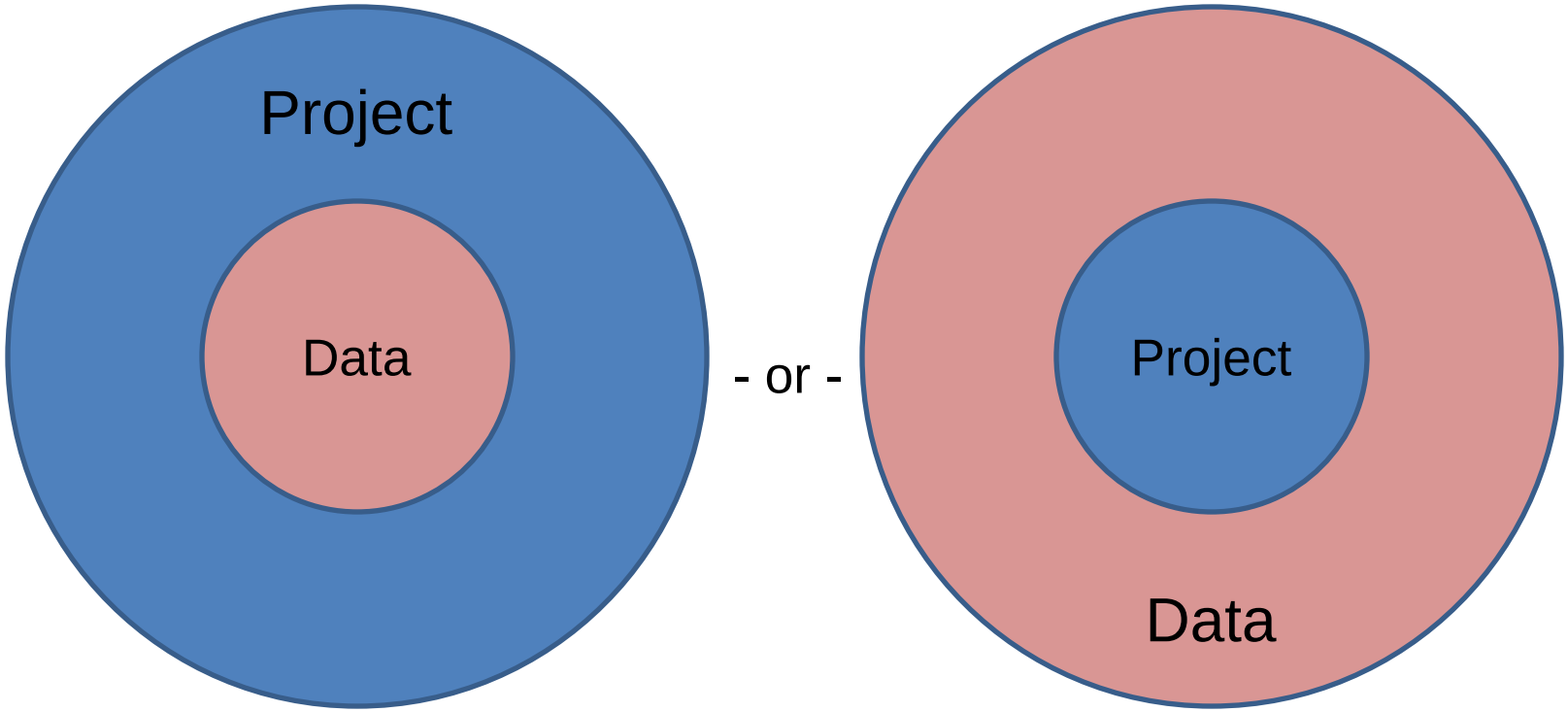
During a Project: DMP becomes an *implementation guide*, is updated to reflect reality, accumulates documentation as it is produced.

End of Project: DMP represents a *complete record of project data activities*, "evidence", and metadata catalog.

2. Data cannot exist without a process behind it

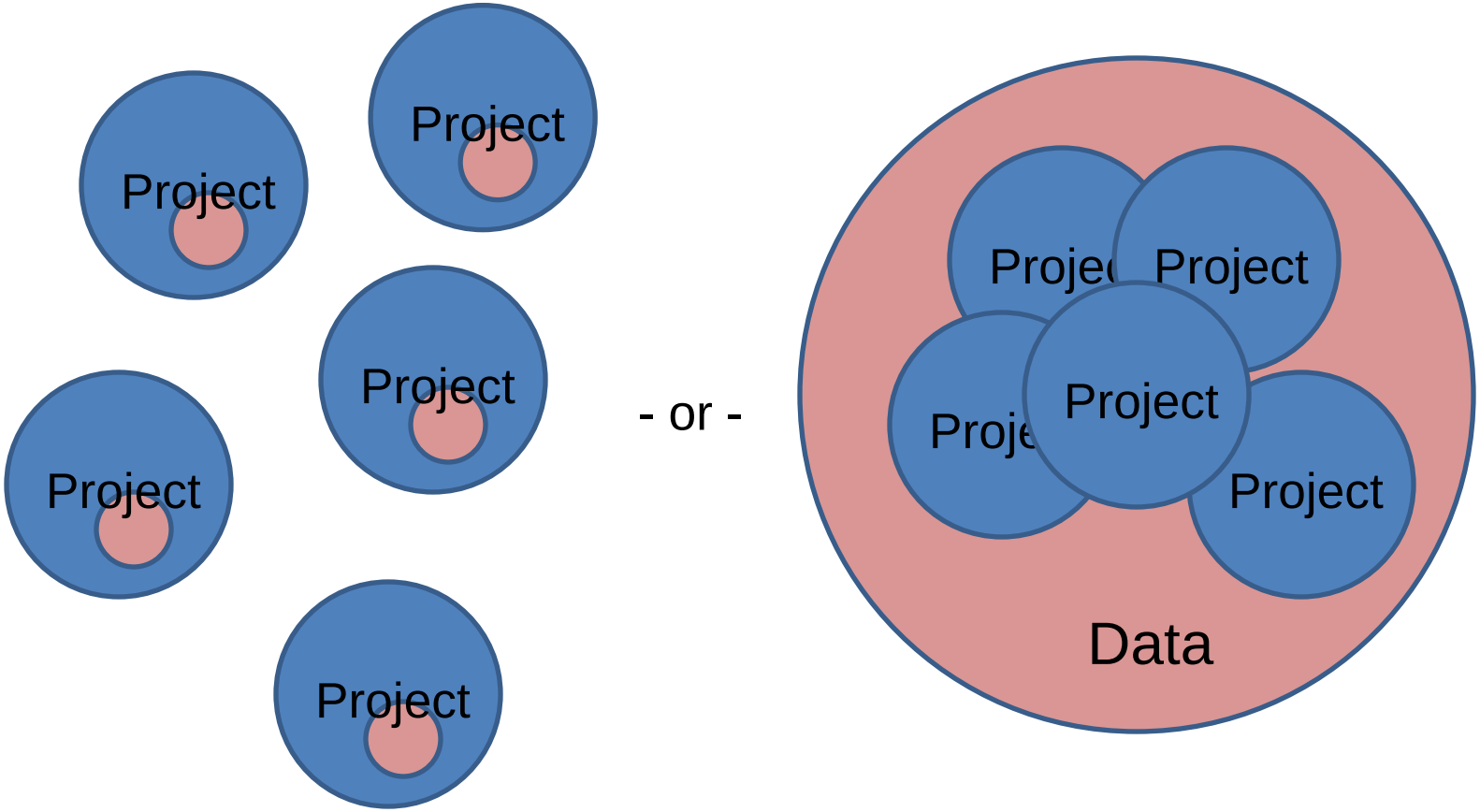
Observation, data collection, existing data acquisition, deriving new value types, integrating disparate data, computational modeling, statistical abstraction, data product development

Eye Test



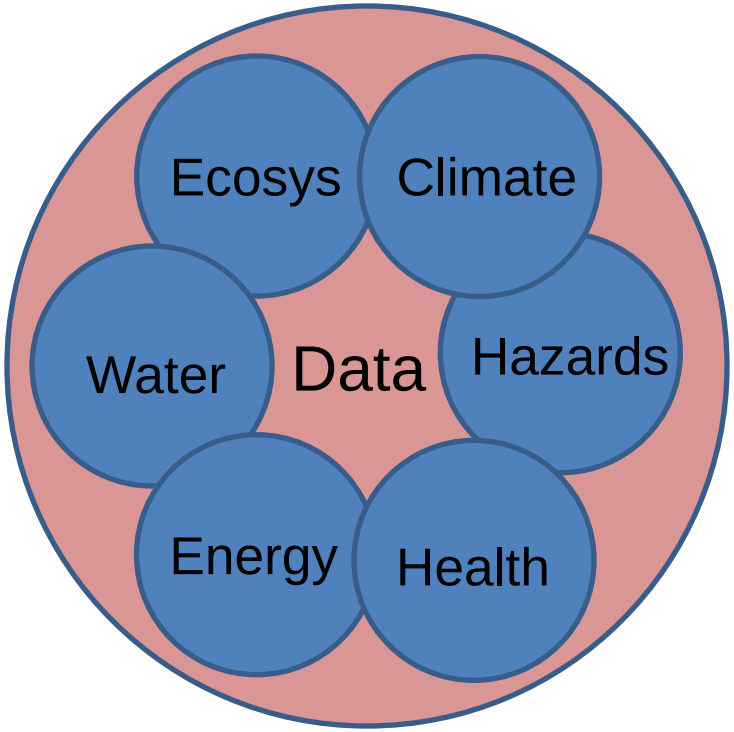
What do you see?

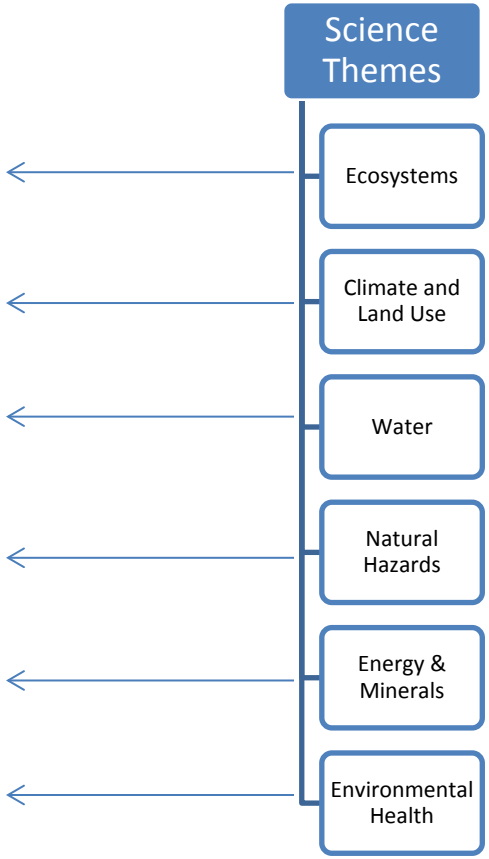
After Many Projects



Integrated Data Resources

Data Management Planning can include Planning for Integration



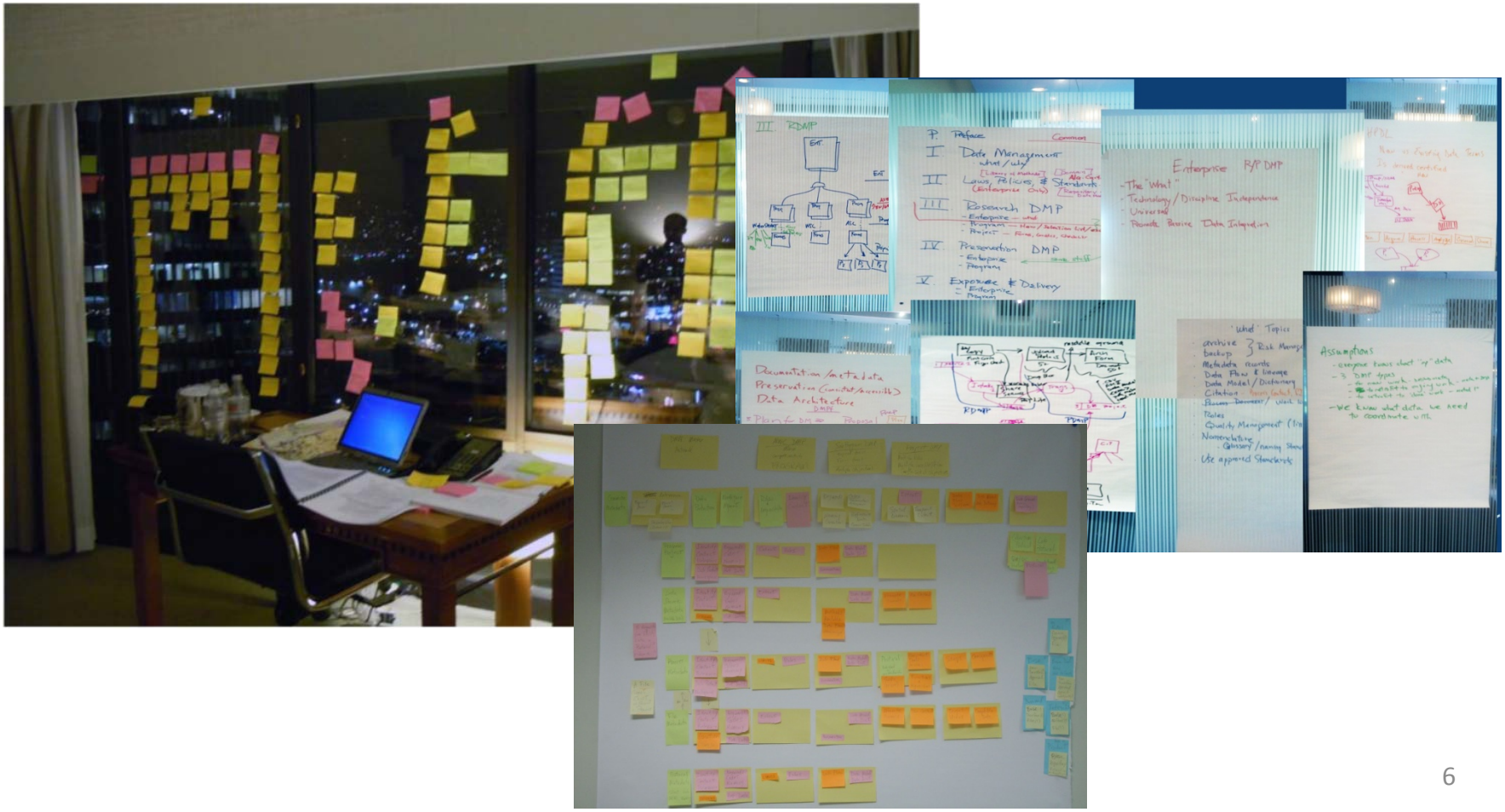


We can get here from there

Data Management Planning Framework (DMPf)

Mike McHale, Stan Smith, Steve Tessler, with Tom Burley joining in later

* Origin was with the Climate Effects Network (CEN) in 2010 and 2011
* Additional development was funded in part by CDI, Texas Water Science Center, National Water Census, and Alaska Science Center



DMPf Core Concepts

1. Split Research off from Preservation

Proposed by UK Data Archive, and others

Message to Researchers - “You don’t have to do it all”

DMPf also splits off Data Exposure and Delivery (sharing)

2. Recognize Data Levels

Used by NASA, NEON, and others

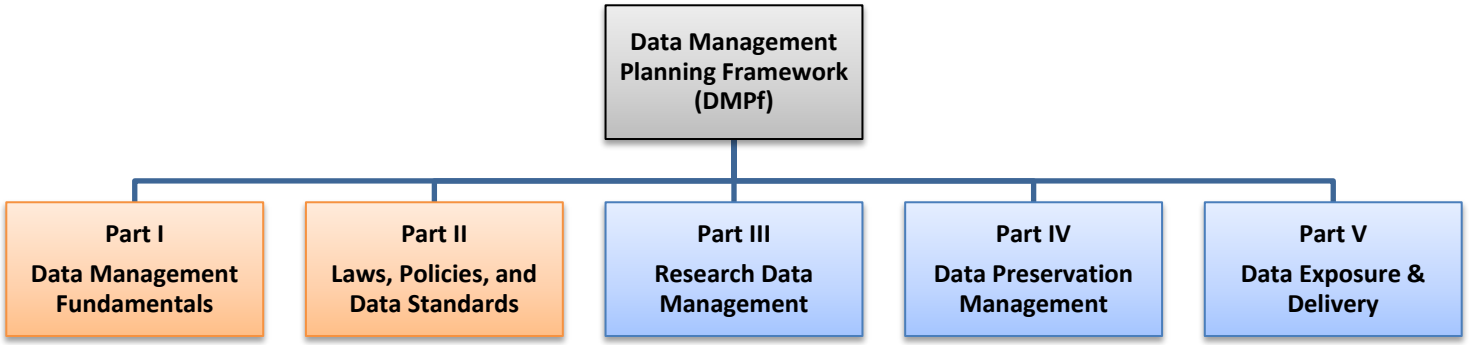
Raw Data, Base Data, Derived Data, Project Database, Data Products

3. Document Layering

Used by NPS Inventory & Monitoring Program

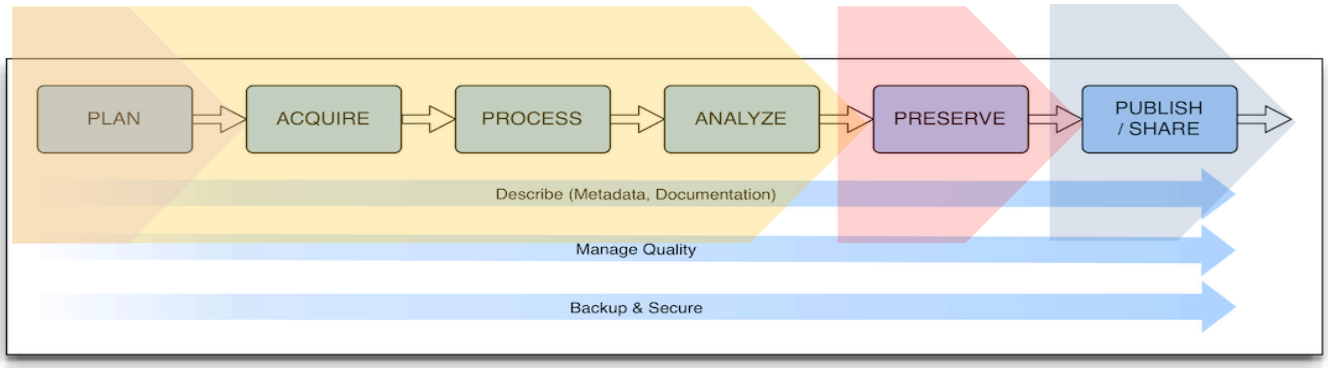
Overall DM Inheritance, Enterprise requirements, Program requirements, Project requirements

DMPf Structure – 5 Parts



Orientation and Supporting Documents

Major Phases of Data Management



Compatible with USGS Science Data Lifecycle Model

DMPf Part III - Document Layering

Data Management

- The things *common to all DMPs*
  - Concepts
  - Roles
  - Vocabulary
  - Glossary
  - …

Enterprise Requirements

- *What was decided*
  - Laws and Policies
  - Adopted standards
  - Standards adoption process
  - Intellectual property policy
- *The 'what' must be done*
  - QA/QC
  - Metadata
  - Backup
  - Alternate storage
  - …

Program Requirements

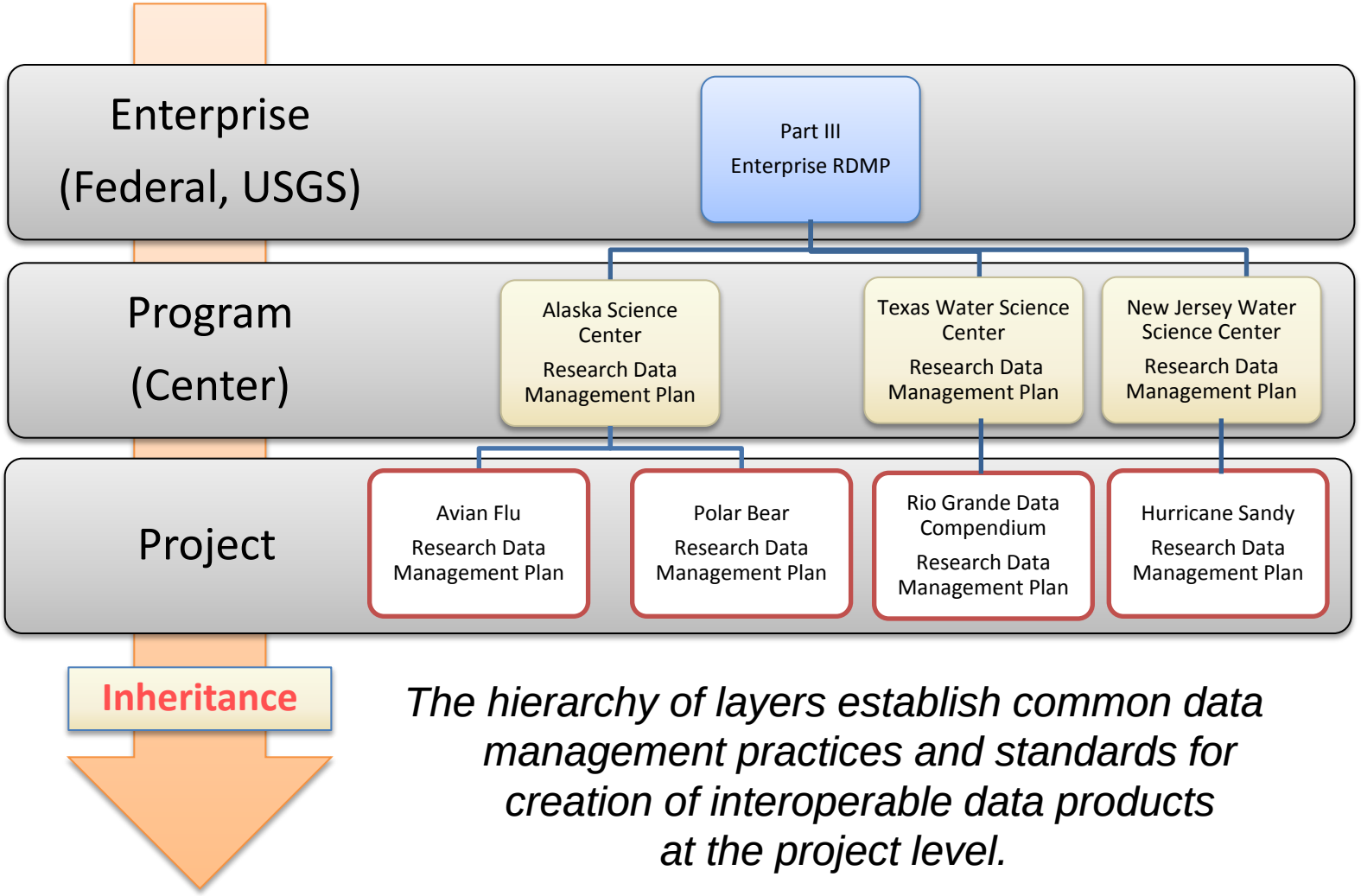
- Program objectives
- *The 'how' things will be done*
  - QA/QC
  - Metadata
  - Protocols
  - …
- Extensions and deviations from Enterprise
- Role Assignments
- Set degree of freedom for the Projects
- …

Project Requirements

- Project objectives
- *The 'details'*
  - Data models (ERD)
  - Metadata
  - Collection forms
  - Extensions and deviations from Program
  - Role Assignments
  - Data transition plans
  - …

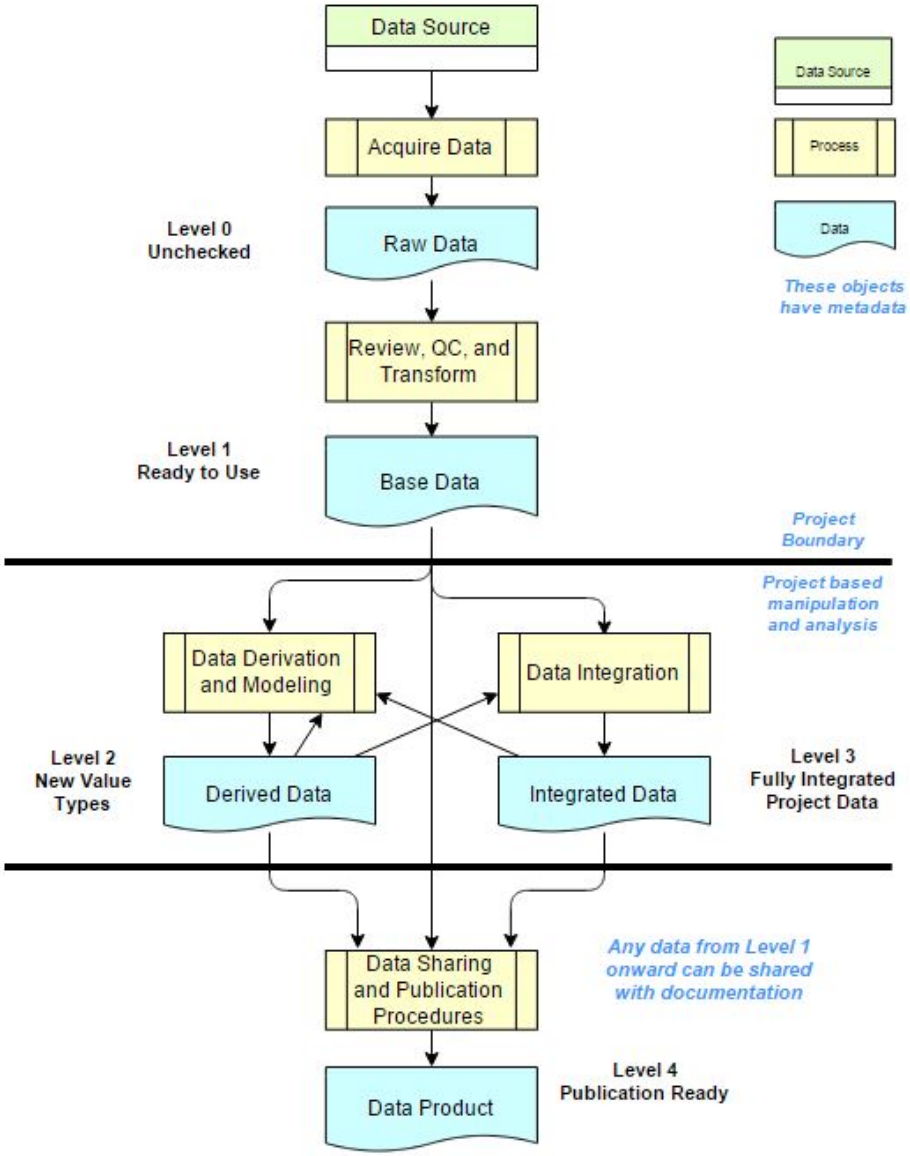
Things to do, what must be done and why, how to do it and why, and what was done.

DMPf Part III - Document Layering



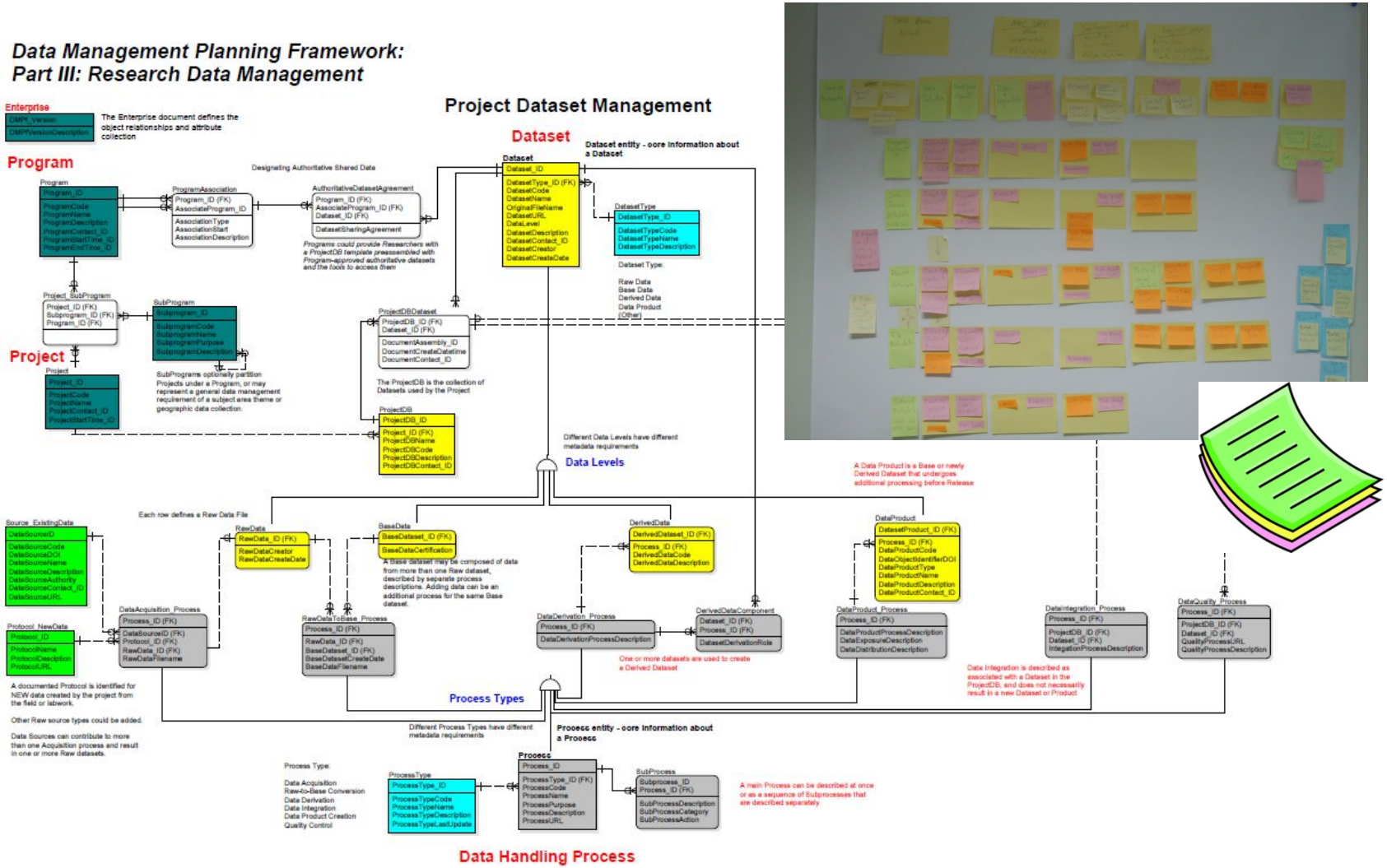
The hierarchy of layers establish common data management practices and standards for creation of interoperable data products at the project level.

DMPf Part III - Data Maturity Model



Every object has metadata

DMPf Part III – Modeling Metadata



Having a good idea is one thing. Selling it is another.

Handing off the whole set of ideas in a single volume has not worked (even though it's exciting reading).

Neither has handing off the whole thing as a collection of individual things (still looks the same to most people).

The concepts make sense, though, so what next?

"In theory there is no difference between theory and practice. In practice there is."

Yogi Berra

So how can we make research data management practical?

Find level of management support

A. Draft a complete DMP for management review
   - Result: no meaningful acceptance
   - Try it again, same results

B. Turn it around – Let management lead …
   - Present management with a menu of data management activities
   - Discuss each activity separately
   - Explain how the activity contributes to the science project and to meeting the needs of management, researchers, and fulfilling high-level requirements
   - Ask what level of support management has for that activity
     - (none, do later, do now)
   - Only include activities in the DMP that management supports

Find level of management support

- Used a simple spreadsheet
- Chose 30 candidate activities from the DMPf
- Organized the DM activities into 5 categories
  - Center/IT responsibilities
  - Project planning
  - Project implementation
  - Project research
  - Project close
- Indicated for each activity
  - Addresses requirements (OSTP/OMB/FSP)
  - Contributes to collecting required metadata
  - Estimates of the implementation effort

Find level of management support

ASC DM Activities 2014-03-31

ID	Current Status	Data Management Activity	Notes	Contributes Metadata	Address OSTP/OMB Req'mnt	Implement Now/Later/Skip
Center/IT						
1	Partial	Project space backup strategy	user recoverable, version history, delete protection, multiple off-site copies		Yes	Now
2		Maintain standard nomenclature list	acronyms / abbreviations		Yes	Skip
3	ASC Portal	"Program" level metadata	narrative, list of projects, intended uses of data		Yes	Now
4		Data Sharing/Management Policy			Yes	Now
5		Data Management Plan/Guidance			Yes	Now
6		Maintain standards for data	internal portal page		Yes	Later
7		Maintain standards for file formats	internal portal page		Yes	Now
Project - Planning						
10	Partial	RDMP project summary	abstract, temporal scope & scale, spatial scope & scale	Yes	Yes	Now
11	Tested	RDMP Conceptual Data Model (CDM)				Later
12		RDMP source datasets	new datasets, existing datasets	Yes	Yes	Now
13		RDMP product datasets		Yes	Yes	Later
14		RDMP responsible parties	by role type	Yes		Now
15		RDMP hardware/software architecture				Now
16		RDMP data management budget			Yes	Later

Set Management objectives for DMP

- To protect our investment in data …
  - Data is expensive to collect and process
  - Earth science data represents a state in nature which cannot be recreated
  - Data has long-term value which often increases over time and outlives current analysis
- To meet our obligation for accepting public money for research as defined by OSTP and OMB
- To make ASC's data available for future projects and scientists
- To highlight the accomplishments of USGS and ASC by making project information and data publicly available

Set DM goals for DMP

- A policy statement that defines Center, PI, and DM responsibilities
- A set of procedures that result in …
  - ❑ Raw and base data being archived
  - ❑ Appropriate data being published
  - ❑ All archived and published data described by metadata
  - ❑ ASC web portal being updated
  - ❑ Secure and regular backups of project folders
  - ❑ A ‘dashboard’ to identify DM status for a project
- We *don’t* want a DMP that just checks the box: ☑ I did it

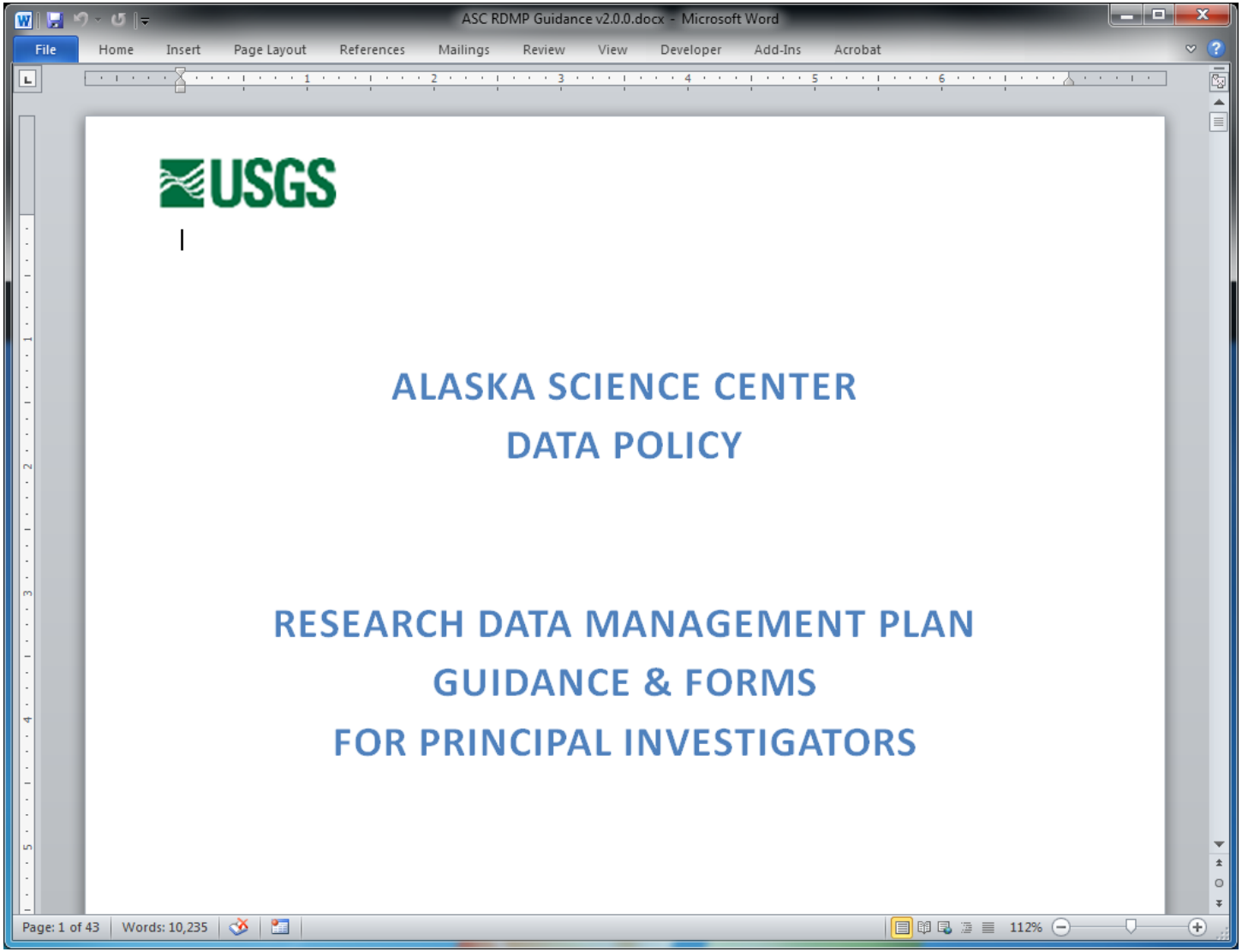
Target the document audience

- (R)esearch Data Management Plan (ASC RDMP)
  - Guidelines and Forms for Principal Investigators
  - Define the data management responsibilities of the research team during the research cycle
  - Interpret inherited responsibilities (from OSTP, OMB, DOI, USGS)

DMPf	III. Research DMP	IV. Preservation DMP	V. Access DMP
National	OSPT, OMB	OSPT, OMB	OSPT, OMB
USGS	FSP, DMPf	FSP	FSP, WRET
Center	**ASC RDMP**	-	-

- OSTP – Office of Science Policy & Technology
- OMB – Office of Management & Budget
- FSP – USGS Fundamental Science Practices
- DMPf – Data Management Planning framework
- WRET – USGS Web Re-Engineering Team

Parameters for DMP have been set



ASC RDMP Table of Contents

I. Objective

II. Data Policy

- Short and sweet policy statements

III. RDMP Process

- Overview of RDMP process by research phase

IV. RDMP Workflows

- Process flow charts for PI to follow

V. RDMP Forms

- Collect all metadata

VI. Glossary

- Terms and roles used in RDMP

VII. Accepted Archival Formats

- Borrowed from NARA and filtered to supportable subset

VIII. References

ASC RDMP Workflow

The ASC RDMP defines a flexible *event-driven, object-oriented* model. Meaning …

- There ***is*** no common or prescribed workflow defined for all projects
- Rather, the ASC RDMP is organized into 12 common project events
- Events can occur in any order, multiple times, or not at all during a project's research cycle
- When one of these events occurs the PI will visit that event's workflow to see their data management responsibility for the event
- Responsibility generally includes completion of one or more ASC RDMP forms

ASC RDMP Events

Data Management Events during research …

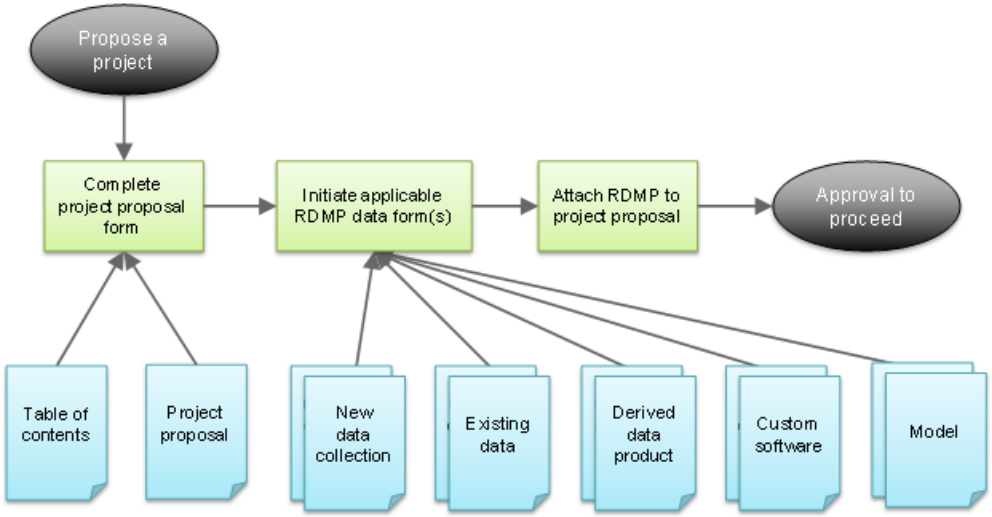
1. Submit a project proposal
2. Project selected for funding
3. Design spreadsheet
4. Design database
5. Build database
6. Design data collection form
7. Archive raw data
8. QC raw data | archive base data
9. Conclude research cycle
10. Package data for distribution
11. Publish data
12. Maintenance of archived & published data

ASC RDMP Event 1

1. SUBMIT PROJECT PROPOSAL

Each new project proposal is required to attach a brief Research Data Management Plan (RDMP). The proposal-level RDMP outlines the project's anticipated data requirements and serves as a skeleton for full RDMP should the project be funded. Details will be added to the RDMP through the research cycle once the project is approved and as research progresses.

a. Complete project proposal form -> The proposal-level RDMP requires a completed Project Proposal Form accompanied by a Table of Contents.
b. Initiate data forms -> Separate New Data Collection, Existing Data, Derived Data Product, Custom Software, and Model form for each data, software, or model item planned for the project. Note data, software, and model forms are divided into sections corresponding to phases of the research project. Only the top section needs to be completed to submit with the project proposal
c. Attach RDMP to project proposal -> Attach the RDMP forms to the project proposal for submission to management for consideration.



ASC RDMP Forms

- Forms are designed primarily to collect ***metadata*** that …
  - ❑ Describe the project
  - ❑ Describe raw and base data
  - ❑ Capture Process/Data Quality/Lineage Steps
  - ❑ Describe data to be published
  - ❑ Support ASC web portal pages
- Forms are currently in MS WORD as they are still in flux
- When stable, we will convert these to web-forms that support a backend database
- Current forms collect ~95% of all metadata needed for constructing robust ISO 19115-2/19110 or FGDC records

ASC RDMP Forms

Research Data Management Forms…

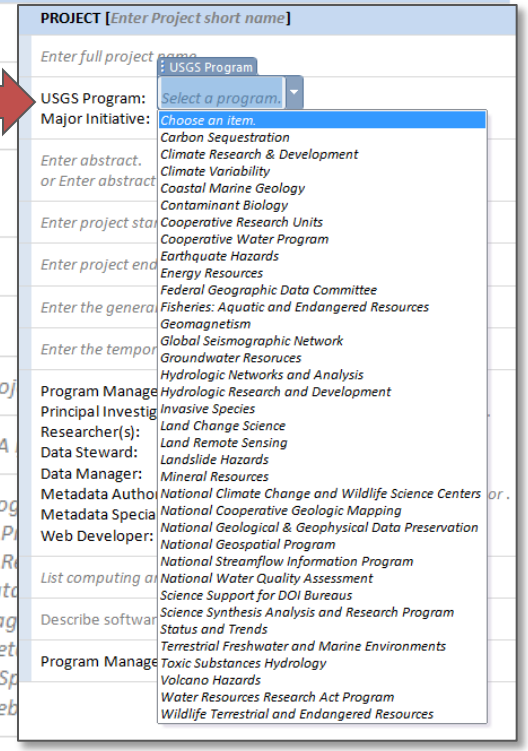
1. Project Proposal Form
2. Project Implementation Form
3. RDMP Forms Table of Contents
4. New Data Collection Form
5. Existing Data Form
6. Derived Data Product Form
7. Custom Software Form
8. Model Form
9. Data Description Form
10. Data Domain Form
11. Process Step/Data Quality/Lineage Form
12. Contact Form

Forms assist completion

ASC PROJECT PROPOSAL FORM

This form is to be completed during the project proposal/planning phase.

	PROJECT [*Enter Project short name*]
Title:	*Enter full project name*
Funding Source:	USGS Program: *Select a program.* Major Initiative: *Select an initiative.*
Abstract:	*Enter abstract.* *or Enter abstract path/file name.*
Project Start Date:	*Enter project start date.*
Project End Date:	*Enter project end date.*
Spatial Description:	*Enter the general geospatial scope of the proj*
Temporal Description:	*Enter the temporal scope of the study or N/A*
Responsible parties:	Program Manager: *Enter name of Prog* Principal Investigator(s): *Enter name(s) of P* Researcher(s): *Enter name(s) of R* Data Steward: *Enter name of Data* Data Manager: *Select Data Manag* Metadata Author: *Enter name of Met* Metadata Specialist: *Select Metadata Sp* Web Developer: *Enter name of Web*
Hardware environment:	*List computing and field hardware requirements.*
Software environment:	Describe software requirements.
ASC manager sign-off:	Program Manager signature/date:



Forms support data maturity model

ASC NEW DATA COLLECTION FORM

Complete this form for data that does not currently exist and will be collected or generated during the project; for example, a new field data collection. This form will follow new data as it matures from raw to base to publication.

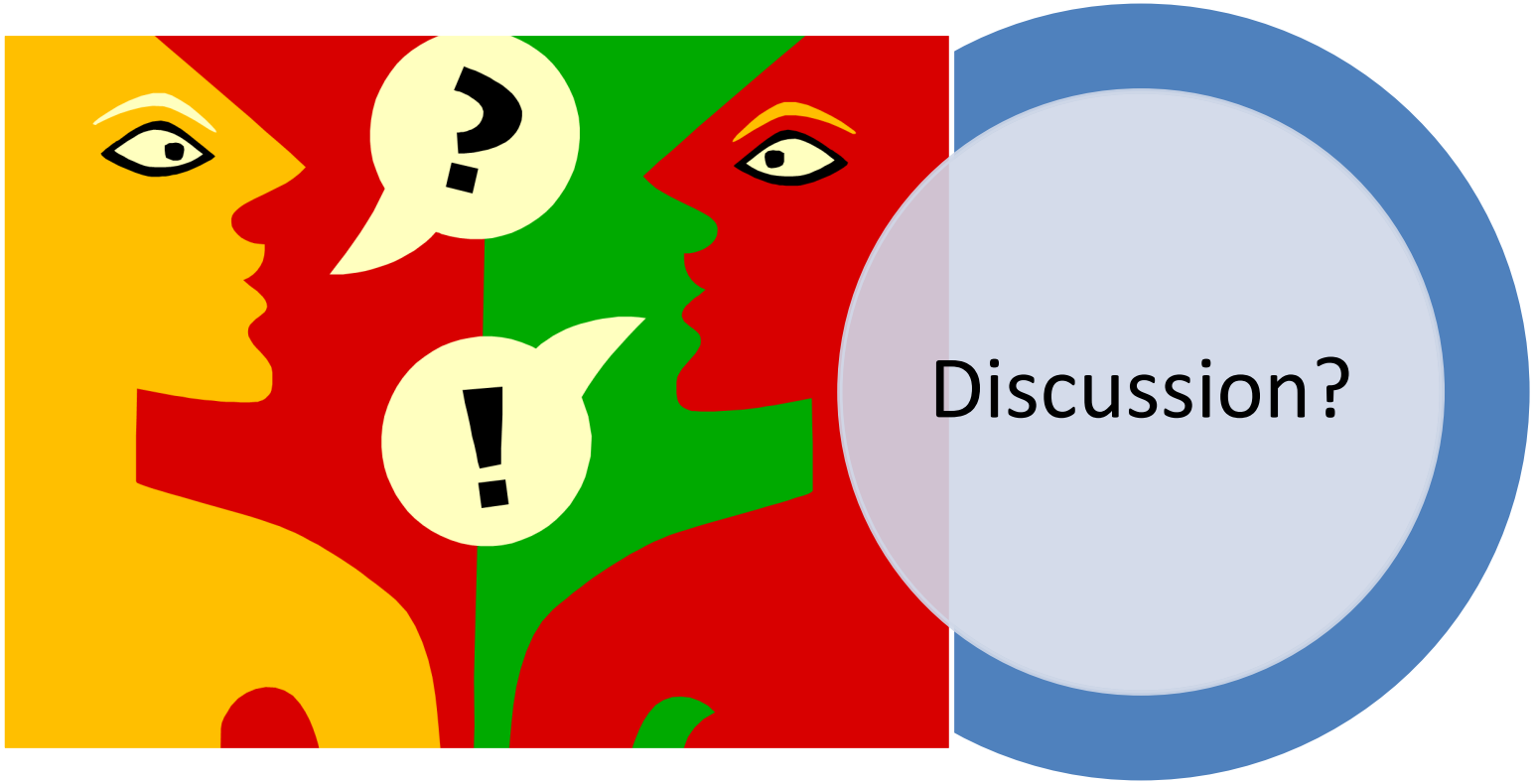
ID [*Enter ID*]	
---------- To be completed for project proposal ----------	
Name:	*Enter new data collection name.*
Description:	*Enter description for new data collection.*
---------- To be completed for raw data ----------	
Format:	*Select format.* *If other, enter format.*
Data coll	
Data coll	

Contact:	Form ID: *Enter Contact Form ID of contact for the raw data.*
---------- To be completed for base data ----------	
Format:	*Select format.* *If other, enter format.*
Data description:	Form ID: *Enter Data Description form ID:*
Data Pr	
Keywo	
Data vo	

Contact:	Form ID: *Enter Contact Form ID of contact for the base data.*
---------- To be completed when data is made public ----------	
Exclusive use period:	Period: *Select period of exclusive use.* Justification: *Enter justification for period of exclusive use.*
Data Distribution Repository(s):	*Enter data repositories other than the Authoritative Repository.*
Citation:	*Enter data citation.*

Set realistic expectations

- This is new - new to management, data managers, PIs, and even a new concept in RDMP design
- We start, find what doesn't work and fix it
- Point out errors and bottle-necks as you encounter them
- Suggest improvements
- Above all, please do not struggle …
  - Contact your Data Manager early on for clarification
  - Rely on your Data Manager to help complete any task



Links:

- Stan Smith – stansmith@usgs.gov, Steve Tessler – stessler@usgs.gov
- Alaska Science Center DMP download from USGS Data Management Website
- Alaska Science Center DMP online version (USGS internal)