Implementing ORD's Scientific Data Management (SDM) Strategy

Lynne Petterson, Ph.D., on detail to ORMA/IMTS
EPA Conference on Managing Environmental Quality Systems
April 23, 2008


Office of Research and Development
Office of Resources Management & Administration (ORMA), Information Management & Technology Staff (IMTS)

Scientific Data Challenges

- Sheer volume & complexity of scientific data
- Research is more collaborative (multi-discipline, multi-location)
- Lack of enterprise-wide definitions for information types & values
- Non-existent or non-consistent file & directory naming schemes
- Scientists must retrieve an entire file to ascertain its relevance
- No access to important metadata in scientists notebooks or heads
- Un-owned data with unknown or dubious content

- **What this means:**
  - Ineffective use of limited budgets
    - Research may be duplicated
    - Potential productivity may be lessened
    - Research may be lost in event of a disaster

2

EPA's Scientific Data Are Important

- "EPA's data resources represent one the Agency's greatest assets. As a national Federal source of reliable and comprehensive statistical information on the state of public health and the environment, EPA is uniquely equipped to provide the public with critical tools to pursue responsible policies."

  Browner and Hansen, EPA Reorganization Memorandum (February 1997)

- "Will someone 20 years from now, not familiar with the data or how they were obtained, be able to find data sets of interest and then fully understand and use the data solely with the aid of the documentation archived with the data set?"

  Committee on Geophysical Data, National Research Council, Solving the Global Change Puzzle (National Academy Press, 1991)

- "Lab personnel spend more than 16 hours a week managing data (and) Lab managers spend up to 8 hours per week trying to find data..." Indiana School of Informatics (2005)

3

Need for ORD-wide Data Strategy

2006: Electronic Laboratory Notebook (ELN) Study (ORD)

2005: IT Improvement Plan (ITIP) Recommendations (ORD)

2005: Federal CIO Council Strategic Plan 2007-2009

2005: Long-Lived Digital Data Collections (Draft, NSF)

2002: EPA Strategic Information Plan (OEI)

2002: ORD's 2002 Information Management (IM) Action Plan (SUSS Consulting)

2002: Major Management Challenges & Program Risks: EPA (GAO)

2002: Scientific Information Management Coordination Board (SIMCorB) Business Plan (ORD)

2002: Barriers to the Effective Management of Government Information on the Internet and Other Electronic Resources (Interagency Committee on Government Information)

1999: Environmental Information: EPA is Taking Steps to Improve Information but Challenges Remain (GAO)

4

1997: ORD's Scientific Information Management Plan

ORD Scientific Data Management (SDM) Strategy
- Comprehensive, Long-Term Approach
- John Sykes led team creating SDM Strategy

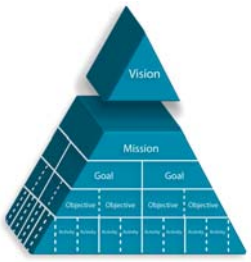
SDM Vision

"ORD will be recognized as a leading federal organization for protecting, managing, and preserving scientific data, and thereby furthering its mission through an integrated framework of knowledge sharing and collaboration."

SDM Mission

"To support the scientific community in fulfilling ORD's mission by providing the policies, standards, practices, and technology necessary for accessing, maintaining, and preserving the ongoing integrity of scientific data."

Goals (5)
Objectives (40)
Tasks (170)



Protect
Maintain
Preserve
Share

5

Achieving the Mission

SDM Mission will be achieved with 5 strategic goals:

- **GOAL 1**: Establish processes that safeguard ORD's scientific data throughout its lifecycle.
- **GOAL 2**: Identify and comply with Federal and Agency requirements for long term preservation and archiving of scientific data.
- **GOAL 3**: Ensure the quality and integrity of ORD scientific data.
- **GOAL 4**: Facilitate collaboration and long term accessibility to ORD's scientific data assets.
- **GOAL 5**: Enable ORD's scientific community to share and manage its data.

Protect – Maintain – Preserve - Share

6

How measure progress?

- **Measure progress against a Capability Maturity Model (CMM):**
  - Point of reference for appraising current SDM processes
  - Framework for prioritizing activities
  - Way to define SDM improvement within a larger context
  - High level snap shot of where SDM is now and where it is going

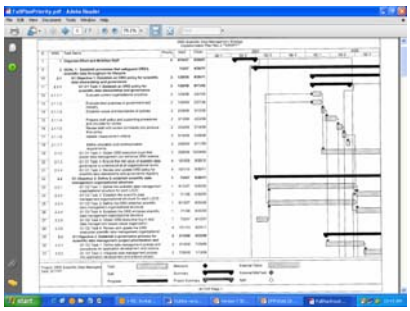
- **CMM Levels:**
  - Level 1: **Random** (SDM processes are ad hoc, rely on individual heroics, unlikely to be able to duplicate any successes)
  - Level 2: **Defined** (Greater attention to documentation and standards)
  - Level 3: **Repeatable** (Project management techniques are used, successes could be repeated)
  - Level 4: **Measurable** (Organization is able to monitor and analyze SDM activities and processes)
  - Level 5: **Optimal** (Processes are continually improved through feedback and innovation)

7

Protect - Maintain – Preserve - Share

Journey of a thousand steps

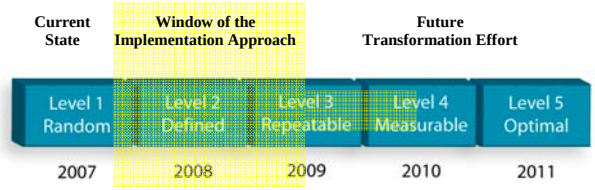
- **Implement SDM strategy with project management tools & techniques**
  - Guide SDM activities from start to finish
  - Document each Goal, Objective, and Task
  - Show associated CMM Levels
  - Provide estimates for start and end dates, with associated Gantt chart
  - Highlight critical dependencies that must be completed before subsequent tasks can be initiated
- **Convene ad hoc Tiger Team to ensure stakeholder input**
- **Propose SDM functions, roles and responsibilities for new IT/IM organizational structure**



8

Near Term Implementation

- **In July 2007, convened a team of ORD Managers at the Branch Chief level or above**
  - Input provided by IT, IM, QA, RM staff and database coordinators
- **Which SDM activities should be tackled over the next 12-18 months?**
  - Contribute to Maturity Levels 2 or 3
  - Define and establish policies, procedures, and practices
  - Have strategic importance or long duration
  - Are critical path activities with dependencies that drive the schedule
  - No technology roll outs



9

Protect - Maintain – Preserve - Share

Specific Near-Term Objectives

Goal 1 (Objective 2): Define & establish scientific data management (SDM) oversight

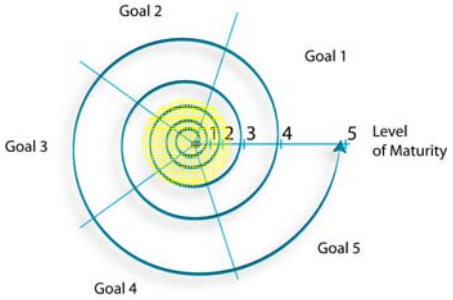
Goal 2 (Objective 1): Define and publicize scientific record classification and retention policies

Goal 2 (Objective 4): Establish standards, policies, and procedures for data preservation, cleanup, change control and audit

Goal 3 (Objective 3): Establish standards, policies, and procedures for scientific data quality cleanup, change control and audit

Goal 4 (Objective 2): Establish scientific metadata standards, policies, and procedures

Goal 4 (Objective 3): Develop scientific e-records ontology and taxonomy



10

Protect - Maintain – Preserve - Share

What does this mean?

- **Propose a framework for SDM oversight**
  - e.g. Scientific data stewardship roles & functions (curators/custodians)
  - Provide guidance and education
- **Revisit scientific record classifications and retention schedules**
  - Enterprise-wide guidance and education
- **Guidance for deleting or changing data**
  - Ensure changes to scientific data are traceable
  - Design change control procedures for modifying or deleting data
- **Standards and guidance for metadata**
  - Use Dublin Core elements (title, creator, subject, date, etc)
  - "Quality" indicators (help users determine the data's objectivity, competence, truthfulness, accuracy, timeliness)
  - "Project" indicators for ORD
- **Develop ORD-wide e-records ontology and taxonomy**

11

Taxonomy/Ontology: *"Cool as a Moose"*

Taxonomy

- Classification of things, as well as the principles underlying the classification
- Set of controlled vocabulary terms, usually hierarchical; Once established, can help inform navigation and search systems
  - New York City
    - *Parks: Central Park*
    - *Area: 843 acres*
    - *Pedestrian paths: 58 miles*
    - *Trees: 26,000*
    - *Benches: 8,968*

Ontology

- Resemble taxonomies but use richer semantic relationships between term and attributes, as well as strict rules about how to specify terms and relations
  - Do more than a control a vocabulary
  - Often thought of as knowledge representation
  - Support browsing and searches; Interoperability for knowledge management

12

Finding ORD data without a taxonomy

13

Finding ORD data with a taxonomy

14

Benefits of SDM Strategy

- **Follows a comprehensive approach, rather than a piecemeal one**
  - Easier access to scientific data
  - Improvements in efficiency and throughput
  - Operational consistency
- **Focuses on benefits to researchers**
  - Increased research collaborations from heightened awareness and visibility
  - Greater insights by internal/external collaborators and stakeholders
  - Ability to identify useful relations (knowledge discovery)
  - Reuse of existing data for new studies
  - Flexible in both long-term and near-term implementation
- **Improves long-term data quality**
  - Follow-up and monitoring

15

How ensure success?

- **Communicate frequently**
  - Clearly publicize purpose to those effected by new guidance
  - Emphasize "WIIFM"
- **Ensure continuing participation by scientific research community, IT/IM, QA, and RM**
- **Tackle small, parallel activities to improve ORD SDM**
  - Emphasize small steps that can be implemented at the team or Branch level that have a big impact over time
- **Remain nimble and responsive to new and evolving ORD SDM priorities**

16

Protect - Maintain – Preserve - Share

EPA United States Environmental Protection Agency

And Then……

- **Get moving**
- **Communicate continuously**
- **Be flexible and responsive**

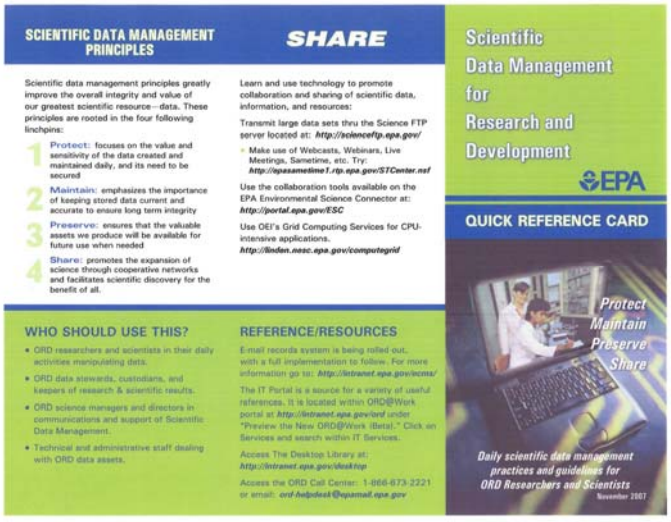
- **Special thank you to Near Term Implementation Team:**

Jerry Blancato	Abdel Kadry	Bhagya Subramanian	Laura Doyle
Christopher Zarba	John G. Lyon	Robert Shepanek	Brenda Culpepper
Myriam Medina-Vera	Steve Young	John Sykes	Deborah Wales
MacArthur Long	Kevin Kirby	Lora Johnson	Jerry Waterman
Steve Jordan	Brenda Young	Margie Vasquez	Jerry Bennett, SRA
Tony Olson	Paul May	Thomas Hughes	
Paul Lemieux	Steve Greenfield	Nancy Broom	
Douglas Young	Linda Harwell	D.B. Ray	
Michael Gonzalez	Ronald Vormwald	Carry Croghan	
Jeff Frithsen	Gary Walter	Valerie Brandon	

17

EPA United States Environmental Protection Agency

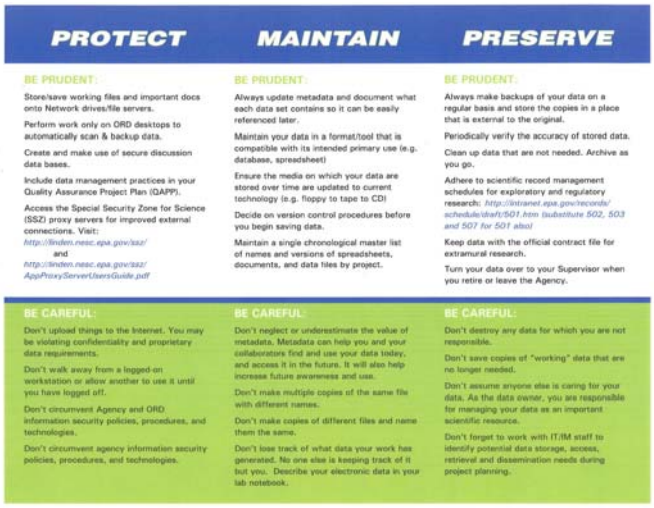
SDM Brochure (front)



18

EPA United States Environmental Protection Agency

SDM Brochure (back)



PROTECT	MAINTAIN	PRESERVE
BE PRUDENT:	BE PRUDENT:	BE PRUDENT:
Store/save working files and important docs onto Network drives/file servers.	Always update metadata and document what each data set contains so it can be easily referenced later.	Always make backups of your data on a regular basis and store the copies in a place that is external to the original.
Perform work only on ORD desktops to automatically scan & backup data.	Maintain your data in a format/tool that is compatible with its intended primary use (e.g. database, spreadsheet)	Periodically verify the accuracy of stored data.
Create and make use of secure discussion data bases.	Ensure the media on which your data are stored over time are updated to current technology (e.g. floppy to tape to CD)	Clean up data that are not needed. Archive as you go.
Include data management practices in your Quality Assurance Project Plan (QAPP).	Decide on version control procedures before you begin saving data.	Adhere to scientific record management schedules for exploratory and regulatory research: http://intranet.epa.gov/records/schedule/draft/501.htm (substitute 502, 503 and 507 for 501 also)
Access the Special Security Zone for Science (SSZ) proxy servers for improved external connections. Visit: http://linlon.nesc.epa.gov/ssz/ and http://linlon.nesc.epa.gov/ssz/AppProxyServerUsersGuide.pdf	Maintain a single chronological master list of names and versions of spreadsheets, documents, and data files by project.	Keep data with the official contract file for extramural research.
		Turn your data over to your Supervisor when you retire or leave the Agency.
BE CAREFUL:	BE CAREFUL:	BE CAREFUL:
Don't upload things to the Internet. You may be violating confidentiality and proprietary data requirements.	Don't neglect or underestimate the value of metadata. Metadata can help you and your collaborators find and use your data today, and access it in the future. It will also help increase future awareness and use.	Don't destroy any data for which you are not responsible.
Don't walk away from a logged-on workstation or allow another to use it until you have logged off.	Don't make multiple copies of the same file with different names.	Don't save copies of "working" data that are no longer needed.
Don't circumvent Agency and ORD information security policies, procedures, and technologies.	Don't make copies of different files and name them the same.	Don't assume anyone else is caring for your data. As the data owner, you are responsible for managing your data as an important scientific resource.
Don't circumvent agency information security policies, procedures, and technologies.	Don't lose track of what data your work has generated. No one else is keeping track of it but you. Describe your electronic data in your lab notebook.	Don't forget to work with IT/IM staff to identify potential data storage, access, retrieval and dissemination needs during project planning.

19

EPA United States Environmental Protection Agency

ORD Taxonomy Next Steps

- **Build ORD Taxonomy from researcher perspective**
  - Assist in accessing, distributing, exchanging, and sharing scientific data & other information
- **Identify agreed-upon scientific terms (controlled vocabulary**)
  - May be used for scientific Records Management, Web taxonomies, information systems such as Environmental Information Management System (EIMS), Science Inventory (SI), and Environmental Science Connector (ESC)

20

ORD Taxonomy Next Steps

- SOW on OEI contract: 18 months with five components
- 1. **General business case**
  - Document costs/benefits
  - Identify/leverage parallel efforts & best practices
  - Identify paper prototype
- 2. **High-level taxonomy based on EPA/ORD strategic mission**
  - 'Loose' hierarchy (organized two or more levels down) & their proposed resolution
  - Tie to relevant metadata schemes (in use or being developed)
  - Reality check use case(s) with target groups

21

ORD Taxonomy Next Steps

- 3. **Mid-term strategic paper**
  - Discuss how ontology/taxonomy would be used (& potential ties to metadata schemes)
  - Document issues associated with integrating information in relevant systems
  - Illustration of how paper prototype would be tested
  - Approach for soliciting feedback from target groups
  - High-level summary for senior management
- 4. **Future maintenance, governance, and stewardship**
  - How terms will be added or deleted
  - Roles and responsibilities
  - Final decision authority
- **5. Final paper**
  - Summary of issues and their resolution
  - Findings from paper prototype testing
  - Putting the taxonomy/ontology to use and extending to deeper levels

22