



Scientific Data Management (SDM) for Government Agencies:  
Report from the Workshop to Improve SDM

# **HARNESSING THE POWER OF DIGITAL DATA: TAKING THE NEXT STEP**

Co-sponsored by the Environmental Protection Agency (EPA),  
CENDI (The Federal STI Managers Group), and the Interagency  
Working Group on Digital Data (IWGDD)  
June 29 – July 1, 2010

Published on March 31, 2011

## LEGAL NOTICES

This report may be used, copied and redistributed subject to the following terms and conditions:

### Citation for Reference or Attribution:

“Harnessing the Power of Digital Data: Taking the Next Step.” Scientific Data Management (SDM) for Government Agencies: Report from the Workshop to Improve SDM. Workshop held June 29 - July 1, 2010, Washington D.C. March 2011. Report No. CENDI/2011-1. Co-sponsored by the Environmental Protection Agency (EPA), CENDI (The Federal STI Managers Group), and the Interagency Working Group on Digital Data (IWGDD)

**Credit line:** "Courtesy CENDI/2011-1" or "Courtesy [Specific Agency or Organization if noted]"

### DISCLAIMER OF ENDORSEMENT

Neither the U.S. Government nor its employees or contractors endorse or recommend any commercial products, processes, or services. Reference to or appearance of any specific commercial products, processes, or services by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the U.S. Government. The views and opinions of authors do not necessarily state or reflect those of the U.S. Government and they may not be used for advertising or product endorsement purposes.

### DISCLAIMER OF LIABILITY

With respect to materials (e.g., documents, photographs, audio recordings, video recordings, tools, data products, or services) neither the U.S. Government nor any of its employees or contractors make any representations or warranties, express, implied, or statutory, as to the validity, accuracy, completeness, or fitness for a particular purpose; nor represent that use would not infringe privately owned rights; nor assume any liability resulting from the use of such materials and shall in no way be liable for any costs, expenses, claims, or demands arising out of the use of such materials.

### Point of Contact

Lynne Petterson  
Office of Science Information Management (OSIM)  
Office of Research and Development (ORD)  
U.S. Environmental Protection Agency  
MD 343-01  
RTP, NC 27711  
Email: [petterson.lynne@epa.gov](mailto:petterson.lynne@epa.gov)

CENDI Secretariat  
104 Union Valley Rd.  
Oak Ridge, TN 37830  
Email: [sdm\\_inquiries@iiaweb.com](mailto:sdm_inquiries@iiaweb.com)

## ACKNOWLEDGEMENTS

Special thanks to the sponsors of this workshop: the IWGDD (Interagency Working Group on Digital Data), the US EPA (United States Environmental Protection Agency), and CENDI (interagency group of senior scientific and technical information [STI] managers from 14 United States federal agencies). We express our appreciation to the sponsors' Workshop Planning Team for their guidance and collaborative support to strategize and organize the workshop. Thanks go to the keynote speakers, EPA's Chief Scientist, the IWGDD Co-chair, our technical plenary speakers, and the discussion panelists for highlighting major scientific data management (SDM) current issues, concerns, and barriers for the federal agencies today. Thanks go to all workshop attendees and breakout group facilitators for their participation and for providing highly valued expertise from a diverse range of SDM fields of practice. Special thanks also go to the report writing team for their perseverance and diligence in developing and critiquing the many drafts of this report.

The organization of the workshop and the production of the report was the result of the dedication of many individuals. Appreciation goes to the following individuals:

### **The Workshop Planning Team**

*Robert Shepanek*, US Environmental Protection Agency, *Chair*  
*Sharon Jordan*, US Department of Energy, *Co-chair*  
*Bonnie C. Carroll*, CENDI Secretariat Support  
*J. R. Candlish*, CENDI Secretariat Support  
*Robert Bohn*, Networking and Information Technology  
Research and Development  
*Stuart Gagnon*, National Agricultural Library  
*Donald Hagen*, National Technical Information Service  
*Lynnann Hitchens*, US Environmental Protection Agency  
*Carol Jacobson*, Defense Technical Information Center  
*Kevin Kirby*, US Environmental Protection Agency  
*Mary Ann Leonard*, National Agricultural Library  
*Deborah Ozga*, National Institutes of Health  
*Lynne Petterson*, US Environmental Protection Agency  
*H. K. "Rama" Ramapriyan*, National Aeronautics and  
Space Administration  
*Sylvia Spengler*, National Science Foundation  
*Bruce Wilson*, Oak Ridge National Laboratory

### **The Report Drafting Team**

*Robert Shepanek*, US Environmental Protection Agency  
*Coordinator*  
*Bonnie C. Carroll*, CENDI Secretariat Support, *Co-Coordinator*  
*J. R. Candlish*, CENDI Secretariat Support, *Co-Coordinator*  
*Jeffery Campbell*, National Oceanic and Atmospheric  
Administration  
*Denise Duncan*, LMI Government Consulting  
*Kevin Kirby*, US Environmental Protection Agency  
*Lynne Petterson*, US Environmental Protection Agency  
*Joanna Pratt*, Stratus Consulting Inc.  
*H. K. "Rama" Ramapriyan*, National Aeronautics and Space  
Administration  
*Holly Surbaugh*, Stratus Consulting Inc.  
*Bruce Wilson*, Oak Ridge National Laboratory

### **Report Compilers:**

*Robert Shepanek*, US Environmental Protection Agency  
*Bonnie C. Carroll*, Information International Associates, Inc.  
*J. R. Candlish*, Information International Associates, Inc.  
*Rose B. Raney* and *Andrea Baumgartner*, Editorial and  
Formatting Support, Information International Associates, Inc.

## TABLE OF CONTENTS

1.0 INTRODUCTION .....	1
1.1 Workshop Background .....	1
1.2 Workshop Objectives.....	1
1.3 Workshop Approach .....	2
1.3.1 Workshop Planning Effort .....	2
1.3.2 Bibliography.....	2
1.3.3 Survey .....	3
1.3.4 Plenary Sessions.....	3
1.3.5 Workshop Breakout Sessions.....	4
1.4 Key Context for Data Policy and Management Plans.....	5
1.4.1 Project Lifecycle.....	6
1.4.2 Data Lifecycle .....	7
1.5 Introduction to Sections 2.0 and 3.0.....	7
2.0 DATA POLICY .....	8
2.1 Statement of Guiding Principles for Digital Scientific Data Preservation and Access ...	8
2.2 Managing Scientific Data as Enterprise Assets or Liabilities .....	9
2.2.1 Portfolio Management.....	10
2.2.2 Culture Issues.....	11
2.3 Scientific Data should be Managed According to an SDM Plan that Covers the Full Data Lifecycle.....	12
2.4 Description of Mechanisms for Access to Specialized Data Policies.....	13
2.5 Retain Data Commensurate with Its Value .....	14
2.6 Ensure that SDM Processes are Integrated with Knowledge Management Initiatives .	15
2.7 Identify Scientific Data with Metadata to Enable Needed Business Operations .....	16
2.8 Manage Scientific Data for Appropriate Control.....	18
2.9 Maintain Version and Change Control on Datasets .....	19
2.10 Assignment of Responsibilities .....	20
2.10.1 Communities of Practice (COPs) .....	21
2.10.2 Reward Structure.....	21
2.11 Statement of Intentions and Mechanisms for Cooperation, Coordination, and Partnerships .....	22
2.12 Provisions for Updating and Revisions .....	22
2.13 Data Management Policy Outline .....	23
3.0 DATA MANAGEMENT PLANS .....	25
3.1 General Considerations for Science Data Management Plans.....	27
3.2 Observations on Specific Data Management Plan Elements .....	29
3.2.1 Description and Impact (30%).....	30
3.2.2 Data Governance (84%) and Stewardship (81%).....	30
3.2.3 Data Sharing (68%), Access (84%), Data Security Management (81%), and Version Control (92%) .....	31
3.2.4 Metadata Management (97%), Content and Format (74%), Document and Content Management (63%) .....	33
3.2.5 Preservation (84%) and Transfer of Responsibility (70%) .....	35
3.2.6 Data Architecture (69%) and Database Operations Management (69%).....	36
3.2.7 Reference and Master Data Management (7%) and Data Warehousing and Business Intelligence (54%).....	37
3.2.8 Data Quality Management (97%), Provenance (88%), and Usability (61%) .....	37
3.2.9 Value-added Services for the Data (30%) and Workflow Systems (35%).....	37

4.0 RECOMMENDATIONS .....38

4.1 Recommendations to Federal Policy .....38

4.1.1 In Order to Better Manage Scientific Data as an Enterprise Asset, Policy and Research Funding Agencies Should Support Gaining a Better Understanding of the Value Proposition for Effective Scientific Data Management .....38

4.1.2 Agencies Should Consider Portfolio Management as a Model when Determining How to Allocate Resources to Manage and Preserve a Complex Array of Data Generated or Held by an Agency .....38

4.1.3 Agencies Should Stimulate Cultural Change through a System of Incentives to the Stakeholders .....39

4.1.4 A Solid Interagency Coordination Function Should be Established and Maintained .....39

4.1.5 A Scientific Data Research Agenda Should Be Established to Provide an Objective Foundation for Scientific Data Management Decisions.....40

4.1.6 Open Government Goals Should be Supported by Data Management Policy and Planning.....40

4.2 Recommendations to Agencies on Data Policy .....40

4.2.1 Each Agency Should Have a Data Policy that Should be Developed in a Federal Policy Context and Should be Compatible with Programmatic and Community of Practice Policies .....40

4.2.2 Agencies Should Manage Scientific Data for Appropriate Control while Making Appropriate Access More Transparent.....41

4.2.3 Agencies Should Establish the Role of Chief Data Officer and should Clarify Roles and Responsibilities of Agency Personnel .....41

4.3 Recommendations to Agencies on Data Management Planning.....42

4.3.1 Scientific Data should be Managed According to an SDM Plan that Covers the Full Data Lifecycle and Also Supports the Full Project Lifecycle.....42

4.3.2 The Data Management Plan Should Be “An Ongoing, Open-ended Living Electronic Record” that Follows the Data through its Lifecycle .....43

4.4 Recommendations on COPs .....43

4.4.1 Agencies Should Make Effective Use of COPs .....43

4.5 Recommendations on Infrastructure .....44

4.5.1 Agency Infrastructure Should Support Scientific Data Management .....44

4.5.2 Policy and Individual Agencies Should Support Persistent Identification across the Government, Including Version Control to Facilitate Data Management and Use .....44

4.5.3 Data Management Planning In and Across Agencies Should Include a Commitment to Effective Metadata Management.....45

5.0 CONCLUSION.....46

APPENDIX A Community of Practice Lifecycle Models..... A-1

APPENDIX B Policy Elements and Principles ..... B-1

APPENDIX C SAMPLE Data Policy for Digital Scientific Data Based on Federal and Industry Best Practices ..... C-1

APPENDIX D PROJECT Data Management PlanTemplate ..... D-1

APPENDIX E Workshop Agenda ..... E-1

## LIST OF FIGURES

1.4-1	EPA project lifecycle .....	5
1.4-2	IWGDD digital data lifecycle model (NSTC, 2009) .....	5
2.0-1	Data policy framework of elements and components .....	8
2.1-1	Guiding principles from the <i>Harnessing Report</i> .....	8
2.2.1-1	Types of data collections .....	10
2.5-1	Perceived decision makers on data retention policy .....	15
2.6-1	Information object linkages to data sets .....	16
2.7-1	Pre-workshop survey rankings of elements of data management plan elements .....	17
2.7-2	Participants' perspectives on the sufficiency of data quality for use .....	18
2.13-1	Popularity of required and optional policy elements requested for policy structures....	23
2.13-2	Seventeen additional policy elements requested by survey participants .....	23
3.1-1	Data management plan elemental matrix composed from three resources: <i>The Harnessing Report, the EPA Survey, and DAMA's Functional Framework</i> .....	26
3.1-2	Hierarchical structure of data management planning functions .....	27
3.1-3	Prioritized data management plan elements most useful in defending policy Recommendations denoted by survey participants .....	28
3.2-1	Most identified barriers for the secondary use of scientific data .....	30
3.2.4-1	The level of documentation needed to validate the quality of "found" data .....	34
3.2.5-1	Process to transfer responsibility for data objects .....	36
A1	The FGDC data lifecycle (FGDC, 2010) .....	A1
A2	A linear data lifecycle for use in an operational environment (EPA, 2010b) .....	A2
B1	Elements of policy and definitions collated from <i>Harnessing Report</i> and the <i>EPA Survey</i> .....	B1
C1	The science project lifecycle .....	C4
C2	The generic scientific data lifecycle.....	C5

## EXECUTIVE SUMMARY

*Empowered by an array of new digital technologies, science in the 21<sup>st</sup> century will be conducted in a fully digital world. In this world, the power of digital information to catalyze progress is limited only the power of the human mind. Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity. (NSTC, 2009)*

As science becomes more data driven, advances in technology provide extraordinary opportunity to find, manage, manipulate and use data and information. To achieve the full promise of these advances for the enterprise of federal science, effective, well-crafted federal scientific data management policies and plans are necessary. The Interagency Working Group on Digital Data (IWGDD), a component of the Office of Science and Technology Policy, recognized the need for attention to scientific data management policies and plans in their report, “Harnessing the Power of Digital Data for Science and Society.” The goal of the “Harnessing Report” was to provide “a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society.” Two key recommendations were as follows:

- *In laying appropriate policy foundations, agencies should consider all components of a comprehensive agency data policy, such as preservation and access guidelines; assignment of responsibilities; information about specialized data policies; provisions for cooperation, coordination and partnerships; and means for updates and revisions.*
- *The components of data management plans should identify the types of data and their expected impact; specify relevant standards; and outline provisions for protection, access, and continuing preservation.*

### Workshop Background

The purpose of the Workshop on Scientific Data Management (SDM) for Government Agencies (June 29–July 1, 2010) was to provide a forum for federal practitioners of SDM to share best practices in data management policy and planning. Practitioners at the workshop included researchers, science managers, policy analysts, operational users, and data managers. This report documents the results of the workshop.

The workshop was structured to take advantage of the depth of scientific data management experience available within the federal agencies. An extensive reading list was provided to participants in advance of the conference. A preconference survey was conducted to explore participant positions on key topics in SDM policy and planning. Plenary speakers at the workshop shared their experiences with data policy and planning in their federal science programs. These speakers were selected to represent a full spectrum of scientific data categories. Categories included remotely sensed data, data from field observations, large- and small-scale laboratory data, model data, and data from publications. A significant part of the program was devoted to structured breakout sessions to allow participants ample time to share their experiences, insights, and expertise in SDM policy and planning.

### Scientific Data Management Policy

A variety of findings and recommendations on data management policy resulted from the workshop deliberations. Two related themes explored in the workshop are (1) the relationship of the data lifecycle to the project lifecycle, and (2) the possible benefits for data valuation from using portfolio management techniques to evaluate potential projects for funding. Federal policy should encourage the use of these two approaches to help ensure (1) that the long-term value of data is considered in initial funding decisions for projects, and (2) that data management proscriptions and approaches do not create an unsustainable burden on individual project resources. Traditionally, discussions about data management have narrowly focused on the lifecycle of the data. This myopic perspective misses the broader context within which data are created, managed, and preserved, and it can lead to inappropriate policy approaches.

A great deal, but not all, scientific data fall into the category of *potential agency and national enterprise assets*. This strongly suggests that policy should encourage techniques to value data for what are often unanticipated secondary uses. This also suggests that the cost of data preservation be distributed beyond the originating project to other entities. Models discussed at the workshop include (but are not limited to) the following:

- Agency institutional repositories
- Deployment of a federal strategy for use of persistent digital object identifiers
- Use of policy to formalize and encourage development of communities of practice (COPs) to meet SDM challenges

Along with a more rational approach for allocation of resources to support SDM, policy should more aggressively encourage and catalyze a shift in the data sharing culture of science. Policy should reward agencies and individuals for stewardship approaches that include appropriate investments in data management that benefit other and future users.

SDM policy is complex. To be successful it must be outcome-based, and it must include suitable, achievable options to address numerous details. The fundamental questions for scientific data management policy makers are as follows:

- What components should policy include?
- How should these components be defined?
- How can success in these areas be measured in a rapidly changing science and technical environment so that policy adjustments are made in a timely way?

For example, as social media penetrates the culture of science and impacts collaborative approaches, SDM policy must accommodate different approaches to security, data rights, and support for organically developing COPs. Policy discussions and plenary presentations at the workshop covered a wide range of these topics, including:

- Guiding principles for digital scientific data preservation and access
- Management of scientific data as enterprise assets or liabilities
- The need to manage scientific data according to a plan that covers the full data lifecycle with regard for the project lifecycle
- Mechanisms for access to specialized data policies
- Ways to ensure that data are retained commensurate with their value
- The integration of data management with knowledge management
- Proper attention to metadata
- Appropriate control of scientific data
- Version and change control of data
- Assignment of responsibilities for management of data
- Mechanism for cooperation, coordination, and partnerships
- Provisions for keeping policy up to date and relevant
- What areas policy should address

In order to develop and deploy a federal SDM policy for improved access and preservation of the nation's wealth of scientific data, information, and tools, appropriate and sometimes innovative methods to address each of the items listed above will be needed.

### **Scientific Data Management Plans**

Many aspects of SDM policy are implemented through *data management plans*. This is a fast-changing environment, with increasing numbers of agencies requiring data management plans. For example, the National Science Foundation recently required all grantees to submit data management plans as part of their grant proposals.



General considerations for SDM plans discussed at the workshop include the following:

- **One size does not fit all.** SDM plans must be adaptable and tailored for a diverse set of SDM implementation scenarios.
- **Data management plans must be living documents.** As project and data lifecycles unfold, data management plans must be revised to reflect changes. Discussions in the workshop suggested that the concept of the data management plan as a document may not be adequate. The idea of a living data management plan may become an application that links project, plans, data, documents and their metadata together incrementally as project and data lifecycles are executed. These innovative approaches should be explored.
- **SDM planning should be encouraged.** This planning may involve all tiers of the organization, including individual projects, programs, agencies, and COP levels.

The SDM planning chapter provides recommended data management plan contents and specifies the issues to be resolved for effective data management planning that can be institutionalized in the federal sphere. Some key elements of data management plans discussed during the workshop include the following:

- |   |  |
|---|--|
| • Data quality management                         | • Reference and master data management           |
| • Provenance                                      | • Version control and change control on datasets |
| • Access  | • Content and format                             |
| • Data security management                        | • Data governance                                |
| • Preservation & implications for secondary usage | • Stewardship                                    |

Each of these elements was considered and discussed.

## Recommendations

Report recommendations are organized according to which organizations can contribute to successful management of our federal data resource. They are highlighted in the following sections.

### Recommendations to federal policy

In order to better manage scientific data as an enterprise asset, policy and research funding agencies should advocate gaining a ***better understanding of the value proposition for good SDM***. Although no systematic approach was identified at the workshop and there are complex variables with cost/benefit analysis, the value of primary and secondary use should be the ultimate concern.

1. Agencies should consider ***portfolio management*** as a model for determining how to allocate resources to manage and preserve a complex array of data generated or held by an agency. Each agency should make it an enterprise mission to manage its portfolio of data assets and should develop goals for returns on investment and risk tolerances during the evaluation process for these managed collections.
2. Agencies should stimulate ***cultural change*** through a system of incentives to stakeholders. SDM policy should motivate agency researchers to move from the ownership mindset of data hoarding to a data sharing approach. OPM and agency reward structures should reflect these data policy objectives. Such objectives should be mandated through modifications in SDM standards for researcher promotions and through establishment of series for data management and curation. Training should target agency personnel resources toward effective data management. If necessary, personnel resources should repurposed to ensure that adequate resources are available to support effective data management.

3. Organizations must work together to establish and maintain a solid ***interagency coordination function*** to advocate an umbrella policy to address detailed SDM policy at agency, program, project, and COP levels. A National Science and Technology Council (NSTC) subcommittee would be an appropriate mechanism to implement this cooperation. Federal coordination must also connect to other sectors and international activities to draw on their progressive resolutions to SDM challenges.
4. A ***scientific data research agenda*** should be established to provide an objective foundation for SDM decisions. Guidance should be developed on how to evaluate and value scientific data for long-term preservation. This guidance will require systematic understanding and methods, and it should be an integral component of a scientific data research agenda.
5. ***Open government goals*** should be supported by data management policy and planning. It is vital that new tools be integrated into the overall federal architecture and the project lifecycle. SDM policy and planning must comply with the federal objectives of transparency and open access to be sustainable and need to be integrated with the business processes of science to support an interoperable federal architecture.

### ***Recommendations to agencies on data policy***

1. Each agency should develop a ***data policy within a federal policy context*** that is compatible with programmatic and COP policies.
  - a. Agency data policy should adopt ***guiding principles*** tied into the federal data policy context that promote data preservation and access.
  - b. Agency data policy should ensure ***SDM processes integrated with knowledge management initiatives***. It should encourage linkage and presentation of data with relevant information to be provided when possible in a knowledge management context.
  - c. ***Agency policy and responsibilities for data retention*** should be clearly defined. Agencies should retain data commensurate with their value and SDM policy needs to be clear regarding how decisions are made and who should make these decisions. Systematic valuation procedures must be established, and data retention should be a standard part of the records management processes.
2. Agencies should ***manage scientific data for appropriate control while ensuring appropriate access***. Policy should clarify and balance data control versus access. This concept is closely tied to the concept of incentive structures that reward agencies and researchers for sharing data.
3. Agencies should establish the roles of ***chief data officer*** and should clarify roles and responsibilities of agency personnel. It is especially important to develop clear, visible assignments, specifically in regard to data stewardship. Ultimately, SDM policy and planning should be overseen to ensure commitment, diligence, and consistency.

### ***Recommendations to agencies on data management planning***

1. Scientific data should be managed according to an SDM plan that covers the ***full data lifecycle and the full project lifecycle***. Since the full data lifecycle will most likely extend beyond a project, planners must develop an effective strategy to preserve the data. Transfer of responsibility must become a part of the cycle management process and must be conducted in a cost-effective manner through a valuation process to determine the appropriate retention time.
2. The data management plan should be an ongoing, open-ended, living electronic record that follows the data through its lifecycle. The ***living data management plan*** should be initiated early in the project planning process and should be sufficiently adaptable to adjust to project changes. Data management plans should be stored within an easily accessible domain so that they can be used as examples.

### ***Recommendations on communities of practice***

1. Agencies should make ***effective use of COPs***. They should encourage and facilitate planning efforts within and across federal agencies, private entities, and academia.
  - a. An interagency coordination group should develop a ***conceptual framework for COPs*** that clarifies the unique aspects of scientific data compared to technical and administrative data. Guidance is also needed on how the SDM planning function will interact with other agency-wide programs such as project management, quality assurance, and enterprise architecture. Existing governance processes like those included in the chief information officer (CIO) function and enterprise architecture need to be better integrated with SDM governance.
  - b. Agencies should support the role of COPs through ***increased outreach*** to help primary data generators identify future users of their data. Factoring secondary users is important in valuation of data for preservation and access.

### ***Recommendations on infrastructure***

1. SDM needs ***support from federal and agency infrastructure***. Institutional repositories, for example, can provide cost effective resources across many projects and researchers.
  - a. Individual agencies policy should support ***persistent identification*** across the government, including version control.
  - b. Data management planning within and across agencies should include a ***commitment to effective metadata management***. This could include the development of government-wide core metadata standards and core taxonomies.

### ***Conclusion***

There is growing acknowledgement of and response to scientific data as a national asset. Since the workshop in June-July 2010, the role of data has been recognized in the *America Competes Act*, in a recent PCAST report on the “Digital Future,” and in the White House memo on “Scientific Integrity.” We hope the recommendations in this report will be useful to our sponsors and the agencies that they represent.



## 1.0 INTRODUCTION

### 1.1 Workshop Background

On June 29, 30 and July 1, 2010, a workshop to explore best practices in federal scientific data management (SDM) was held in Washington, DC. Eighty-seven federal personnel and other invited contractors participated in the workshop. A total of twenty-five federal agencies and contracting firms were represented. The workshop was sponsored by the Environmental Protection Agency (EPA), CENDI, (the Federal Interagency Scientific and Technical Information [STI] Managers Group), and the National Science and Technology Council (NSTC) Interagency Working Group on Digital Data (IWGDD). Each organization represents stakeholders in SDM with varied but converging interests and perspectives.

EPA's mission to protect human health and the environment depends on the ability to manage scientific data and information within the agency and in collaboration with other organizations. EPA has begun a major initiative to identify best practices as part of its goal to develop agency policies and operating guidelines. This workshop was based on an original EPA proposal to CENDI and the IWGDD and is intended as a major source of input to their initiative.

CENDI is an interagency working group of senior STI managers working cooperatively to improve the productivity of federal science- and technology-based programs through effective STI management. Historically, CENDI's initial focus was on publications, but it has broadened in concert with evolving technology to encompass all digital data objects associated with the scientific endeavor.

The Interagency Working Group on Digital Data (IWGDD) was established by the NSTC's Committee on Science in December of 2006. Nearly 30 agencies, offices, and councils were named as members or participants, reflecting the broad range of interests in digital scientific data. The IWGDD's purpose is to "develop and promote the implementation of a strategic plan for the federal government to cultivate an open, interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering."

### 1.2 Workshop Objectives

In January of 2009, the IWGDD published a report entitled "Harnessing the Power of Digital Data for Science and Society."<sup>1</sup> The report "provides a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society." The SDM workshop focused primarily on two of three recommendations from the report:

- *In laying appropriate policy foundations, agencies should consider all components of a comprehensive agency data policy, such as preservation and access guidelines; assignment of responsibilities; information about specialized data policies; provisions for cooperation, coordination and partnerships; and means for updates and revisions.*
- *The components of data management plans should identify the types of data and their expected impact; specify relevant standards; and outline provisions for protection, access, and continuing preservation.*

---

<sup>1</sup> "Harnessing the Power of Digital Data for Science and Society." Report of the Interagency working Group on Digital Data to the Committee on Science of the National Science and Technology Council, January 2009.

The *Harnessing Report* addresses SDM issues at a policy and strategic level. The goals of the workshop were to assemble federal practitioners to share best practices of their agencies and communities of practice (COP) in SDM, and to gather their ideas for SDM policy and planning to address opportunities to implement the *Harnessing Report* recommendations.

### 1.3 Workshop Approach

#### 1.3.1 Workshop Planning Effort

The workshop was planned and supported by representatives from the sponsoring organizations. This team developed the workshop agenda to explore the practice of federal SDM from multiple perspectives. The team developed the workshop agenda and a pre-conference reading list, and they conducted extensive survey of participants' SDM perspectives. They also developed a list of best practice speakers with expertise relevant to conference objectives. They established a conference structure that balanced examples of agency best practices outlined by plenary speakers with breakout sessions for attendee participation. The planning team's pre-conference efforts ensured a productive workshop. A copy of the agenda is provided in Appendix E. The agenda with links to presentation materials can be found at this [link](#).

#### 1.3.2 Bibliography

To ensure that workshop participants had a common understanding of SDM and to stimulate breakout group discussions, the planning team compiled and distributed background materials to participants. This included a bibliography of existing policies, plans, and key articles on SDM policy and planning. The bibliography was built based on the work of the IWGDD, EPA research, and contributions by the planning team. Three key documents were identified as primary sources of current information on SDM:

#### 1. Harnessing the Power of Digital Data

National Science and Technology Council (NSTC), Office of Science and Technology Policy.

Networking and Information Technology Research and Development (NITRD) Program. 2009.

Harnessing the power of digital data for science and society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council.

Available: [http://www.nitrd.gov/about/harnessing\\_power\\_web.pdf](http://www.nitrd.gov/about/harnessing_power_web.pdf)

#### Getting Started on an Agency SDM Policy: EPA's ORD is Developing SDM Policy and Guidance

In 2007, the EPA's Office of Science and Information Management (OSIM) in the Office of Research and Development (ORD) began to develop a scientific data management strategy and framework to govern data created and used across ORD by its scientists and by ORD contractors, grantees, and other partners. As a first step, ORD developed an SDM strategy, based on its vision that: "ORD will be recognized as a leading federal organization in data management, thereby furthering its mission through an integrated framework of knowledge sharing and collaboration."

ORD then conducted research on how EPA and other federal agencies with missions similar to EPA's are managing their scientific data (see "Survey of EPA and Other Federal Agency Scientific Data Management Policies and Guidance," 2009). While this study identified a wide variety of documents and resources about SDM-related goals, policies, and guidance, it also demonstrated that the federal agencies examined have not yet implemented comprehensive policies and approaches for managing the burgeoning amount of scientific data that they create. Nevertheless, this compilation of resources provided a solid base of information for beginning to develop an ORD SDM policy and related guidance.

OSIM produced a policy statement and SDM framework consisting of eight SDM procedures that are critical to supporting management of the entire data lifecycle. Progress continues on this project, and planned next steps include obtaining management approval for the proposed SDM policy and then issuing implementation guidance with an emphasis on using input and real-world insight from ORD quality assurance staff, scientists, research managers, and other staff at multiple levels. A full rollout of the new framework is currently scheduled for December 2011.

Source: Scientific Data Management Project Summary, EPA Office of Research and Development, 2010

## 2. Survey of SDM Policies

U.S. EPA. 2010. Survey of EPA and other federal agency scientific data management policies and guidance. U.S. Environmental Protection Agency, Office of Research and Development (ORD), Office of Science Information Management (OSIM). Contract No: EPA-600-R-10-04. Available: [http://oaspub.epa.gov/eims/eimscomm.getfile?p\\_download\\_id=496103](http://oaspub.epa.gov/eims/eimscomm.getfile?p_download_id=496103)

## 3. DAMA-DMBOK

Mark Mosley. 2008. DAMA-DMBOK functional framework. Version 3.02. The Data Management Association (DAMA) International. Available: <http://www.dama.org/i4a/forms/form.cfm?id=29>

The *Harnessing Report* represents the IWGDD's work to establish the framework for agencies to take the next step in developing plans and policies. The *EPA Survey* represents the results of one agency's in-depth exploration of the best practices currently in use. These first two primarily represent the mission perspective. The *DAMA Report* delves into defining data management structure and data management functions, and it guides initiatives for implementation. The planning team determined that, taken together, these documents offer a good starting point on which to build. Participants were encouraged to read these documents prior to the conference. The full bibliography is available at this [link](#). These three references were also used to develop suggested best practice elements for a benchmark data management policy and a data management plan. Appendix B, "Policy Elements and Principles" includes the policy matrix; Section 3.0 of this report includes the data management plan matrix.

### 1.3.3 Survey

The planning team determined that a background survey on SDM topics gathering initial ideas from attendees would help provide informed starting points for discussion in the break-outs. An extensive survey of attendees was administered to identify their views, knowledge, attitudes, and use of SDM best practices. Survey questions were grouped into the following broad categories:

- Valuation of data for long-term preservation versus cost of data documentation and storage
- The nature and content of data management policies and plans
- The extent and content of metadata required to support reuse of data
- Appropriate control of access to data
- The need for version control of data
- Data retention based on its value
- Provision of context to support secondary use of data.

The survey population responses stratifies into 4 researchers, 6 science managers, 6 policy analysts, 2 operational users, 25 data managers, and 3 others, with total participation of 46 (a little over 50% of the workshop participants). A summary of the survey results was presented in the opening plenary session. The full results of the survey can be found at this [link](#).

### 1.3.4 Plenary Sessions

Participants were welcomed to the workshop by the EPA Chief Scientist, who described the benefits of EPA's science mission through implementation of sound data management policies and planning. The following keynote address was delivered by Dr. Chris Greer, co-chair of the IWGDD and co-author of the *Harnessing Report*. The keynote address synthesized many of the report's themes and emphasized the IWGDD strategy, to "create a comprehensive framework of transparent, evolvable, and extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data" (NSTC, 2009).

Strategy themes were covered by examples of agency practices in the plenary presentations that followed. Presentations were made by representatives of the National Science Foundation (NSF) the National

Aeronautics and Space Administration (NASA), the Department of Interior (DOI), the National Institutes of Health (NIH), the National Institute of Standards and Technology (NIST), and the National Oceanic and Atmospheric Administration (NOAA). The planning team selected presenting agencies based on their expertise in managing data in one of the following categories:

- Remotely sensed data (satellite, airborne, ground-based, etc.)
- Field data (geospatial, observed, sampled, surveyed, electronic field notebooks)
- Large-scale laboratory data (observed, clinical, sampled, instrument, analytical results, sample handling/tracking, electronic laboratory notebook)
- Small-scale laboratory
- Model data
- R&D publications and their relationships to data

Within these categories, the array of topics presented by the agencies included repository approaches, data management policies, data management plans, governance, developing and fostering communities of practice (COPs), architectures, and a variety of other related information. Copies of the workshop presentations are available at the CENDI website ([link](#)). The workshop concluded with a summary discussion that provided an opportunity for sponsoring organization representatives to respond to what they had learned as conference participants.

### **1.3.5 Workshop Breakout Sessions**

Breakout sessions stratified participants into stakeholder groups whose perspectives on SDM policy and planning transcend agency affiliation. The planning team identified five categories of users:

- Researchers
- Science managers
- Policy analysts
- Operational users
- Data managers

Participants were asked to self-identify as members of one of the five groups. Groups were formed corresponding to the first four categories on the list. The remaining data managers were distributed among the four groups, reflecting the planning team's viewpoint that data managers would provide and gain insight from participation with other categories of users. Of the 87 registered participants, 8 identified as researchers, 18 as science managers, 13 as policy analysts, 2 as operational users, 39 as data managers, and 7 as other.

There were two breakout sessions, each lasting approximately three hours each. The first session was devoted to SDM policy, and the second to SDM planning. The four groups met in separate rooms for these breakout sessions. A plenary session was then held to allow the groups to report their results.

Each breakout session focused on four SDM policy and planning concepts: the project lifecycle, the data lifecycle, proposed SDM policy elements, and proposed SDM plan elements. They also referenced the preliminary background information and plenary session presentations. The breakout sessions allowed all participants to share their perspectives and experience in meeting the challenges inherent in SDM policy and planning.

At the end of each breakout session, facilitators presented the group's findings. These reports can be accessed at the [SDM policy](#) and [SDM plans](#) links.

### 1.4 Key Context for Data Policy and Management Plans

This section introduces key SDM themes from the workshop which are discussed in detail in sections 2, “Scientific Data Management Policy,” and 3, “Scientific Data Management Plans.”

Many different types of digital objects (including the actual data and accompanying metadata, plans, models, and documents) are part of the body of scientific data and the tools of science. The digital era has blurred and extended the boundaries of what constitutes scientific data.

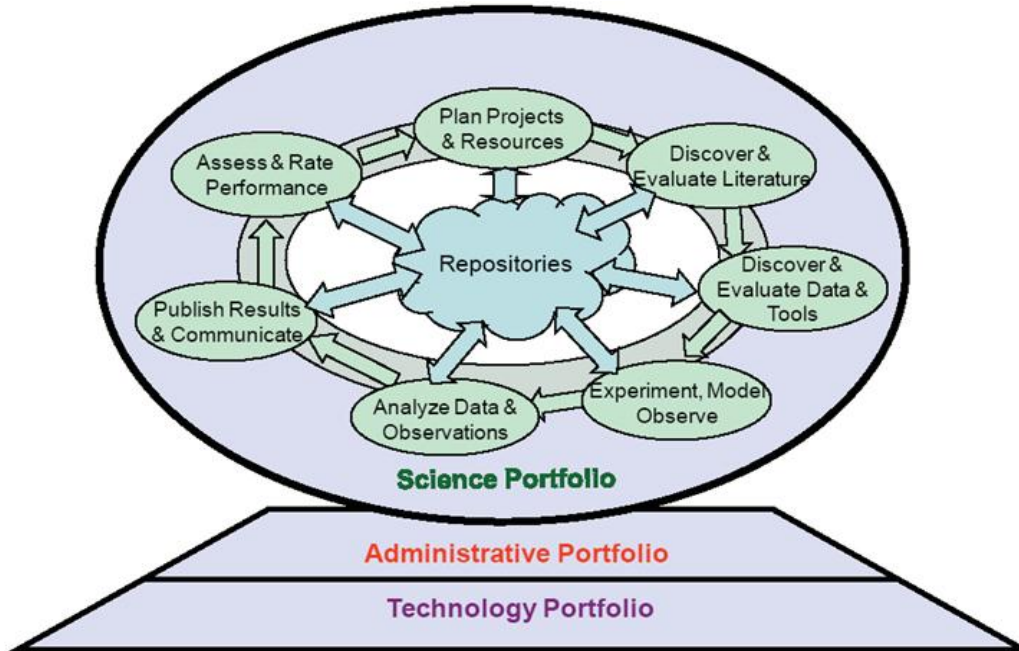


Figure 1.4-1. EPA project lifecycle.

Effective standards and techniques can be implemented to document the broad array of digital objects used in the scientific enterprise. Data management approaches can be developed to maintain relationships among digital objects. Analysis of relationships between different types of digital objects indicates that they must be managed in the context of the *data lifecycle*, which is interdependent with the *project lifecycle*. Workshop survey results and breakout group discussions clearly indicate the need to understand and document data within the context of these two lifecycles.

The project and data lifecycles, as initially presented to breakout participants, are illustrated in figures 1.4-1, “EPA project lifecycle,” and 1.4-2, “IWGDD digital data lifecycle model.” Integration of these cycles was determined to be a best



Figure 1.4-2. IWGDD digital data lifecycle model (NSTC, 2009).



practice, but the integration was only partially implemented in the *Harnessing Report*. This information was provided to breakout groups as background information. These figures show the context of the SDM process, as well as and the breadth of digital objects that actually comprise scientific data. Appendix A provides examples of other COP lifecycles.

The project lifecycle presented in figure 1.4-1 is a generic model of how science is conducted at its most elemental level. Questions are posed, and projects are planned and resourced to answer those questions. Previous results, data, and publications are reviewed for relevance. Experiments are designed and conducted, and results analyzed and published. The cycle repeats, incorporating lessons learned from previous iterations. Because data often have utility beyond individual projects, the data lifecycle (also presented in figure 1.4-2), overlaps but differs from the project lifecycle. Observations on the lifecycles are provided in the two following sections. There are various models for both project and data lifecycles. Those presented herein were used as a starting point for workshop discussions.

### 1.4.1 Project Lifecycle

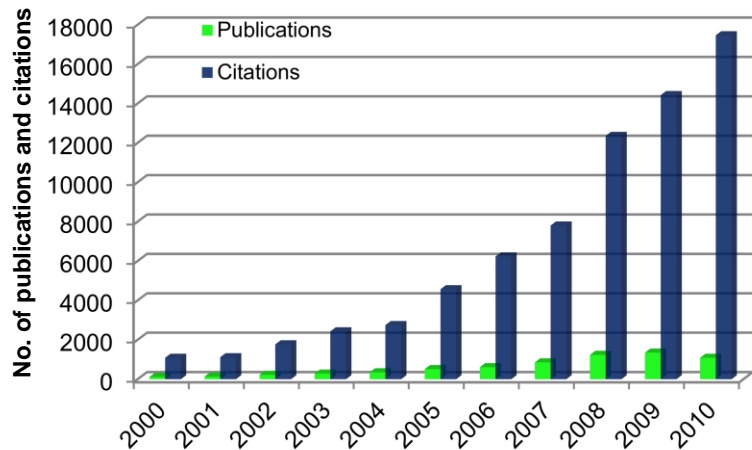
The depiction of the project lifecycle shows the themes pertinent to SDM. Survey results indicate that potential data users want knowledge of data to be generated by new projects and programs, and they also want to be able to understand data in the context of ongoing activities. Some observations are as follows:

- Project descriptions and their relationships to data, tools, publications, and the information they use and produce are important digital objects that need to be managed as part of the data lifecycle.
- Before information and data are produced, an initial, implicit determination of their value is made by the choice to invest in a project.

### NASA's Earth Observing System (EOS) Data and Information System (EOSDIS)

EOSDIS is a geographically distributed system consisting of 12 science data centers, each of which archives and distributes data from a set of specialized Earth science disciplines. With very few exceptions, data held at the EOSDIS Data Centers are distributed openly and free of charge in accordance with NASA's "Earth Science Data Policy." EOSDIS currently holds over 4.5 petabytes of data.

### Publications Resulting from EOS Terra (12/99 launch) and EOS Aqua (05/02 launch) Instruments and Data



### Key Themes

- The publications and citations shown here are a good indicator of scientific growth resulting from NASA's Terra and Aqua missions, which are part of the EOS Program.
- Pre-launch publications and citations are significant, but dramatic growth is seen post launch.
- NASA's EOSDIS, through its well-established data management practices, accomplishes the following:
  - Produces and stores data and metadata in formats compliant with well-documented standards.
  - Provides data, metadata, and software tools promptly to a broad scientific community.
  - Illustrates how data management is a key element in supporting scientific growth

Sources: Terra metrics from Imhoff (2011), Aqua metrics from Parkinson (2011).

- Portfolio management of projects can aid in data valuation for anticipated secondary uses within the enterprise. Various types of data need to be managed to help determine what will bring the best return in the future.
- Descriptions of projects and tools are an integral part of data and document pedigree.
- Effective execution of the project lifecycle depends not only on the architecture of science systems and data, but also on the administrative and technology architecture within an organization.
- A federal architecture that supports interoperable repositories is an important component to providing interoperability within and among projects.
- To be sustainable, SDM approaches must be integrated with the project lifecycle and must impose minimal burden on project resources.

### 1.4.2 Data Lifecycle

The depiction of data lifecycle provided as a starting point for breakout groups was published by the IWGDD in the *Harnessing Report*. Participants were aware of several representations of the data lifecycle and the overlap among them and the project lifecycle. Following are key themes from the discussion of the data lifecycle:

- Culture as a potential barrier or asset to data sharing and secondary use of data
- The need to focus attention on decision making regarding retention, preservation, and transfer of data
- Development of human resources to have the skills and understanding to perform SDM effectively
- The need for organizational structures to be sufficiently agile within and across agencies to address science challenges while also formal enough to be effective in SDM and to address governance and responsibility
- The understanding that data management plans may vary for each SDM environment (e.g., an archive may address ingest of data instead of creation of data)
- Effective SDM to support the broader science enterprise, requiring implementation of an effective policy framework

## 1.5 Introduction to Sections 2.0 and 3.0

The findings and recommendations of this report reflect the viewpoints expressed during workshop breakout and wrap-up sessions, as well as materials provided to participants. Section 2.0, “Data Policy,” proposes a framework for interagency policy to address (1) elements that should be in a policy, and (2) policy considerations connected to best practices discussed by workshop participants. They are organized and framed within the context and best practices summarized in the *Harnessing Report* and the *EPA Survey*. Section 3.0, “Data Management Plans,” specifies the contents that should be included in a data management plan and address issues that need to be resolved for data management planning to be effective and institutionalized in the federal sphere.

## 2.0 DATA POLICY

The *Harnessing Report* suggests five components that should be included in an agency’s scientific data policy. The *EPA Report* suggests development of an SDM policy framework incorporating eight general policy areas. Workshop breakout groups reviewed these policy components and areas, which were presented in a matrix entitled “Policy Elements and Components.” The matrix, provided in Appendix B, includes definitions of each policy area. The charge to the policy breakout groups was to determine which elements should be included in a policy, how that could be applied across agencies, and issues associated with developing guidance for an agency policy. The NASA “Data and Information Policy” and the EPA “National Geospatial Data Policy” are two archetypes, and additional examples can be found within the bibliography ([link](#)).

Based on the output of the workshop, the items in the data policy matrix were reordered as presented in Figure 2.0-1 to provide a logical framework. The following sections address a policy framework in terms of the 12 elements and components.

- 1) Statement of guiding principles for digital scientific data preservation and access
- 2) Management of scientific data as enterprise assets or liabilities
- 3) Development of a SDM plan that covers the full data lifecycle
- 4) Description of mechanisms for access to specialized data policies
- 5) Retention of data commensurate with their value
- 6) Integration of SDM processes with knowledge management initiatives
- 7) Identification of scientific data with metadata to enable needed business operations
- 8) Management of scientific data for appropriate control
- 9) Maintenance of version and change control of data sets
- 10) Assignment of responsibilities
- 11) Statement of intention and mechanisms for cooperation, coordination, and partnership
- 12) Provisions for updating and revisions

**Figure 2.0-1. Data policy framework of elements and components.**

- 1) Science is global and thrives in the digital dimension.
- 2) Digital scientific data are national and global assets.
- 3) Not all digital scientific data need to be preserved.
- 4) Not all preserved data need to be preserved indefinitely.
- 5) COPs are an essential feature of the digital landscape.
- 6) Preservation of digital scientific data is both a government and private sector responsibility, and it benefits society as a whole.
- 7) Long-term preservation, access, and interoperability require management of the full data lifecycle.
- 8) Dynamic strategies are required.

**Figure 2.1-1. Guiding principles from the *Harnessing Report*.**

### 2.1 Statement of Guiding Principles for Digital Scientific Data Preservation and Access

In the pre-workshop survey, 97% of participants supported including of a statement of guiding principles in an agency-level data management policy. The *Harnessing Report* outlined seven guiding principles, as shown in Figure 2.1-1. Workshop discussions and materials support these principles from an operations perspective and provide additional insight into their implications and impact.

Principles 2 and 6 correspond with the elements of policy framework items 2 and 3 shown in Figure 2.0-1. These principles are discussed in Sections 2.2 and 2.3. Relevant support for the other principles is provided throughout the policy discussion.

## 2.2 Managing Scientific Data as Enterprise Assets or Liabilities

The principle that “digital scientific data are national and global assets” is basic to the *Harnessing Report*. This assertion was discussed extensively in workshop breakout groups. The workshop participant survey indicates that over 64% believe that their agencies manage data as enterprise assets. The discussion addressed definitions and culture associated with the data-as-assets view (see Section 2.2.2), as well as other aspects of what it means to view and manage data as assets.

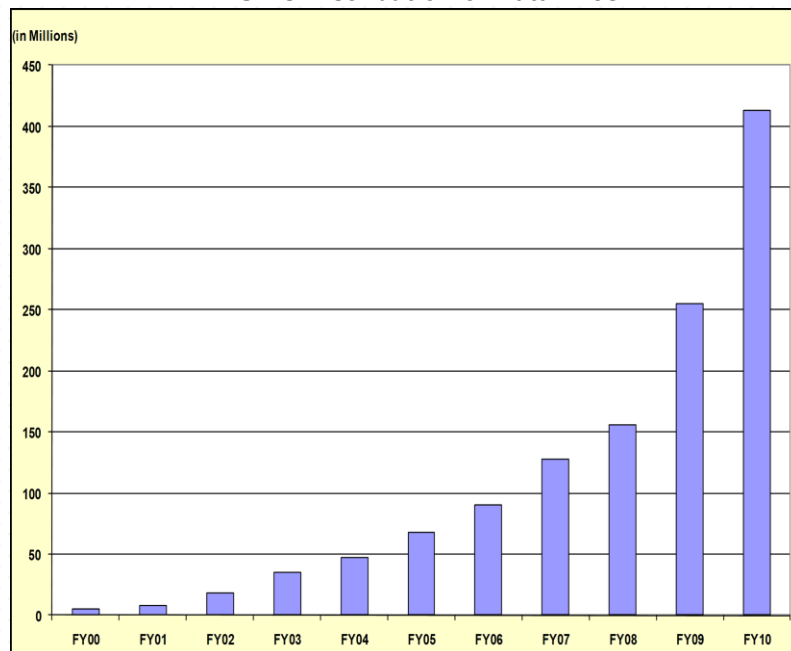
The discussions made it clear that participants viewed scientific data not strictly in the balance sheet context of assets and liabilities, but rather as a resource with intrinsic value. Value calculations are important, but they remain difficult to define. Policy needs to create an environment in which the value of data to the broader scientific community is routinely assessed as part of data management planning. Policy also should include methods to incentivize and compensate data producers for documenting and sharing data. The assessment should address the value of data based on their impact on science and a realistic estimate of the cost of maintaining those data.

### Creating a Culture That Embraces Data Sharing

Workshop attendees generally agreed that open sharing of scientific data has in many benefits. However, it is essential for organizations to (1) acknowledge the potential impacts that such a substantial culture shift might entail for their staff, (2) engage staff in the change process wherever possible, and (3) highlight the positive aspects of the change. One concern researchers have expressed is that data sharing would increase their risk in terms of the use of data as intellectual property. However, several studies illustrate the potential incentives for sharing data. In a study comparing the citation history of cancer microarray clinical trial publications, trials that offered publicly available data were cited by other researchers at a rate 69% higher than trials in which data were not shared.

Source: Piwowar et al., 2007

### NASA's Distribution of Data Files



Distribution of data files of scientific products from NASA's Earth-observing missions has consistently and rapidly grown over the last 11 years due to a combination of open data policy, increased data availability, and systems to support data access and distribution.

Source: NASA EOSDIS, 2010

The analysis of data costs and benefits is complex because it must take into account the overhead costs of maintaining an overall policy and planning framework, as well as the costs of the asset management. This includes such costs as maintaining metadata and the preservation and curation of data. No systematic methodology was identified as a best practice in assessing either costs or benefits. However, case studies, or “vignettes,” were cited that illustrate the impacts of treating data as an asset. The vignettes demonstrate clearly, albeit subjectively, the benefits of sound practices for science and the economy. However, because vignettes are anecdotal and case specific, a more objective, quantitative approach is needed.

Discussions on the value of best data management practices raised the issue of distribution of costs and benefit among data producers and curators, as well as data consumers. This becomes complicated when the producer reaps the initial benefits from the data, such as a scientist getting credit for first publication, yet the data continues to have value for other secondary users. Secondary uses of data can add significant value, as well as cost, to the equation. Participants agreed that federal personnel have a fiduciary responsibility to manage data assets in a responsible manner and that the full data lifecycle should be considered in data valuation. Particular issues arise when data are to be maintained beyond the life of an individual project. Policy and planning approaches are needed to enable complete valuation of data assets, accounting for primary and potential secondary uses.

**The Value of Data Sharing**

In 2003, NIH, the Food and Drug Administration, and numerous researchers from universities, non-profit groups, and pharmaceutical and medical imaging companies joined the Alzheimer’s Disease Neuro-imaging Initiative (ADNI), an unprecedented, large-scale collaboration to locate biological markers showing the progression of the degenerative condition in the human brain. All ADNI participants agreed at the outset to immediately make all findings publicly available, and this openness has resulted in more than 3,200 downloads of the massive data set generated in the first six years of the project with almost a million downloads of the data sets containing brain scan images. And the surge in publication on the topic of early diagnosis of Alzheimer’s could potentially result in promising new treatments.

Source: Kolata et al., 2003

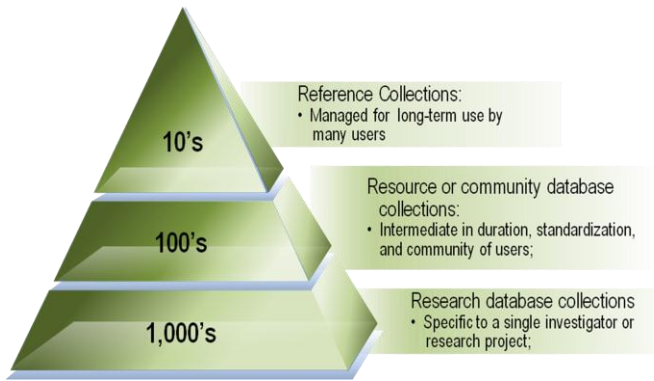
**2.2.1 Portfolio Management**

The portfolio management approach was offered as a model to consider when deciding how to allocate resources to manage and preserve a complex array of data generated or held by an agency. The concept is based on balancing a number of factors to create a strong mixture of assets. Policy should encourage evaluation of data as projects are submitted for initial investment decisions. As in the management of any asset portfolio, each agency must develop its goals for return on investment and for risk tolerance.

One example of the use of portfolio management is the capital planning and investment control (CPIC) process applied to federal information technology (IT) investments. CPIC demonstrates the potential of portfolio management techniques to better align IT investment with agency missions. Similar approaches have been used at EPA in the area of high performance computing (HPC). Finite time available on EPA’s high performance computing platforms is allocated through a competitive review of project proposals and

The phrase “data collections” refers to a dynamic, heterogeneous community system.

- Reference collections: global
- Resource collections: community level
- Research collections: project level



***Informed policy should recognize and build on the existing structure of the DC universe.***

Source: NSB, 2005

**Figure 2.2.1-1. Types of data collections.**

plans. Applying EPA's HPC approach to other scientific computing portfolios would eventually provide a holistic view of EPA's investment in scientific computing.

Agencies participating in CPIC can also assess the value of system data assets and can determine whether management of data generated or used by systems has been adequately planned. Present and future value of data to be produced by science projects should be evaluated as part of the review process. Agency data collections should be reviewed periodically to ensure that resources allocated to preservation are applied for optimum effect. These techniques help evaluators understand the value of individual data sets, as well as their value as part of a managed collection.

Portfolio management groups projects with similar characteristics, or it groups those that are designed to make similar contributions to agency goals and objects. In order to ensure a balanced mixture of solid performing assets and higher risk assets, data can be categorized as presented in the context in Figure 2.2.1-1, "Types of data collections." This figure (from the National Science Board report of 2005 on long-lived data<sup>2</sup>) states that there is a broad base of scientific data from research and development—**research data**—that are useful for the generating project, yet return on investment is higher risk when considering value of the data to the larger community or to secondary users. Some data sets move to a higher level of use—**resource data**—and then a smaller number of data sets move to become generally accepted as valuable **reference data**. Understanding which of the three types of data assets the agency has and then determining which sets belong to each group are an important steps in portfolio management.

EPA applies the concept of **master data**, which is defined as "those data sets and/or elements that are used by more than one program within a segment, programs across segments, or across agencies." Master data are managed by EPA programs/segments answerable to a governing council of users, the EPA Quality Information Council. Understanding the level of community use or the potential use of the broad array of agency data is important in managing this asset.

## 2.2.2 Culture Issues

A majority of survey respondents indicated that they are responsible for managing data as assets. However, with over \$150 billion per year spent on research and development, it is a serious issue that 36% of responses indicate that data (which is a primary output of that R&D investment) are not managed as enterprise assets.

A profound change in the SDM culture must be realized to enable management of data as a national asset. Survey results reveal that resources and culture are considered the two key impediments to managing agencies' scientific data. Combined, over 78% considered resources and/or culture to be the primary impediments to data sharing. Pressure to cause a shift in culture must be applied from the top, while attitudes toward data sharing and reuse within working communities of practice evolve to acknowledge that data are a national asset. Applying pressure from the top requires the formal process of creating policies and plans.

The concept of ownership of scientific data is changing drastically. Research teams need to move from the perception of being **owners** of the data generated by their research to being **custodians** or **stewards** of the data. This approach may engender the perception of losing control of "their" data while being burdened with extra work to serve the needs of others. This shift will be a high-impact cultural change for agency researchers. To effect cultural change, governance should provide incentives for data sharing; and these incentives should be included in the data management plan (See Section 2.10.2).

---

<sup>2</sup> National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21<sup>st</sup> Century," September 2005 ([www.nsf.gov/nsb](http://www.nsf.gov/nsb))

The current federal SDM environment is a patchwork of over-arching and specific data management policies. Some attention has been given to data rights in the same scattered fashion. This approach has resulted in inconsistencies and inefficiencies in preparing data for users outside the original project team. Understandably, without the ability to envision the needs of secondary users, researchers can hardly be faulted for not providing metadata for discovery of the data set beyond their own agency, peer group, or discipline. Increased outreach efforts, support to communities, and development of methods to evaluate data for secondary uses will help primary data generators identify future users of their data.

### 2.3 Scientific Data should be Managed According to an SDM Plan that Covers the Full Data Lifecycle

The *Harnessing* guiding principle—“longer preservation, access, and interoperability require management of the full data lifecycle”—was clearly supported by the planning team’s observation that both data and project lifecycles are critical. Policy should require that data management planning be well integrated into project planning, as noted in Section 1.4.1. The planning should begin at the inception of the project/effort and should be an integral part of project planning, budgeting, and management. The survey and workshop confirmed this need. In fact, 90% of survey participants declared that after a project begins, a data management plan should be “an ongoing, open-ended, living document that follows the data through its lifecycle.” Discussions also strongly supported the idea that a data management plan is a living document. It is this document and its metadata that connect and document the data lifecycle with the project lifecycle. Survey responses indicate that project data and metadata are both important for context to allow effective secondary use of data by operational and policy users.

Federal science agencies are heterogeneous with respect to their scientific data policy requirements and approaches. Data producers often have little incentive to expend resources on extensive data planning that will primarily benefit secondary users. Research agencies vary from regulatory agencies in this regard. Regulatory agencies use their data to develop regulations, so they must prepare the data for a known secondary use. Legal and scientific defensibility must also be addressed in agency policy.

As agencies increase data management requirements, they must also help create or support the infrastructure needed to allow project managers to fulfill policy requirements. This will ultimately change the entire agency culture and practice regarding scientific data management. For example, if an agency provides an institutional repository to its COPs either directly or through supporting initiatives, then multiple projects could incorporate the repository into their plans for storage and archiving. This will help

#### Requirement for a Data Management Plan: NSF’s Response

In the past, scientists at the National Science Foundation (NSF) could expect to maintain records of their research findings in hand-written lab notebooks. But modern technology continues to advance, resulting in explosive growth in data. Now thousands of collaborators across the globe can cooperate on running a single simulation that generates peta-exabytes of data.

In response to this “big bang” of information, NSF has implemented numerous changes related to SDM. NSF will soon augment its long-standing policy on data sharing with new mandates on scientists seeking funding: a two-page data management plan (subject to the peer review process) will be required with the proposal as a criterion for any grant awards. Because data issues resonate with the most strength across NSF units, one important change has been the creation of the Data Working Group, an NSF-wide group of program directors charged with assuring that data are effectively shared within and across disciplines. In addition, in order to encourage data reuse in innovative ways and in combinations not envisioned by those who created the data, NSF has established a Community-based Data Interoperability Networks (INTEROP) program, which funds projects that establish networks to support a wide range of research subjects. Finally, NSF has invested in the Sustainable Digital Data Preservation and Access Network Partners program (called “DataNet”) to develop a widely accessible network of interactive data archives.

Source: Workshop presentation, “NSF Perspective on the Data Deluge,” Philip Bogden, 6/29/2010.

the project manager to define this part of the lifecycle and will help to realize economies of scale and coordination.

As the government looks to its plans for open government through the development of tools such as Data.gov, it is important to integrate these tools into the overall federal architecture and project lifecycle. Federal objectives for transparency and open access to data can only be met sustainably and economically if they are (1) integrated into the business process of science and (2) supported by an interoperable federal architecture. SDM policies and planning are needed to enable this environment to exist.

At the bottom line, agency policy should acknowledge the importance of the project lifecycle context in the data management lifecycle to facilitate data reuse.

## 2.4 Description of Mechanisms for Access to Specialized Data Policies

*Agencies may support various communities of practice and distinct data types, formats, and contexts, and they may have differing programmatic goals, needs, and resources. These agencies should have a harmonized suite of corresponding, specialized data policies. The comprehensive agency digital data policy should describe mechanisms to provide easy and transparent access to the agency's full portfolio of specialized data policies. (NSTC, 2009)*

Workshop participants recognized that science challenges do not respect agency, domain, or discipline boundaries and that scientific data management is the critical capability needed to enable collaboration. They also recognized that interagency policy provides an “umbrella” that must be supplemented by more detailed policy at agency, program, project, and COP levels. Furthermore, these policies should be readily available and should allow the necessary agility for SDM to meet the needs of responders to cross boundary science challenges.

Because COPs often cross agency lines, there must be policies must address interagency consideration. SDM policy must be developed to catalyze COP activities across agencies and to enable collaboration and data sharing, but without stifling innovative approaches at lower levels. Standardized interagency templates, taxonomies, and repositories to manage science data should be considered, and a forum such as the IWGDD provides an excellent vehicle for such considerations. Two key points of context were identified at the workshop:

1. The higher the level of policy (especially at the full agency or interagency level), the more it should be focused on outcomes to allow tailoring to the organization and scientific context. The more that a policy focuses on operations-level activities such as divisions or programs, the more it can focus on process. However, dictation of data management operations is not recommended in policy and should be left to data management planning.
2. Policy in and of itself does not create cultural change. It is a complex mix of factors (including implementation of appropriate rewards structures) to ultimately result in the effective policy implementation.

Workshop participants recognized that COPs and agencies vary widely in their level of sophistication and their commitment to open sharing and systematic management of scientific data. The conclusions of the workshop acknowledge the premise that “one size will not easily fit all” in either policies or plans, but that policy can be developed at an appropriate, generalized level to allow for effective implementation government-wide. Effective policy can contribute to a federal environment where data are viewed as a national asset at all levels.

This conclusion was also borne out in the results of the survey, in which over 94% of respondents agreed that SDM policy could be crafted at a level that could work for their entire agency. Participants were aware that an agency's umbrella policy would have to be supplemented in agencies with domain,



programmatic, and COP policies that would need to be upwardly compatible with it. The *Harnessing* principle —“COPs are an essential feature of the digital landscape”<sup>3</sup>—was confirmed.

The Federal Library and Information Center Community (FLICC) has a long history in information management and increasing involvement in management of many types of digital objects.<sup>4</sup> Many approaches that have been used to manage publications are relevant to SDM. These include the concepts of documenting data, cataloguing data (now elaborated to metadata creation), maintenance of repositories, data archiving, and data reference services provided by qualified personnel. The workshop found that many of these approaches are relevant to those currently needed in data management and should be explored for use.

There are also data management communities in the context of technical data management and in the CIO context. The characteristics of the SDM COP differ from all of these; SDM COP practitioners can learn and borrow from the other COPs. The team drafting this report has relied heavily on input from the CIO data management community and has drawn from the DAMA report.

To facilitate creation of data management plans at all levels, agencies should provide templates or establish an application environment to support planning efforts. Agencies should encourage and facilitate planning efforts not only within their organizations’ structures, but also for COPs which span across federal agencies, other government entities, non-governmental organizations, and private entities.

## 2.5 Retain Data Commensurate with Its Value

Participants expressed complete agreement with the idea that not all data need to be kept or kept indefinitely. They felt that policy needs to be clear about how retention decisions are made and who should make these decisions. Currently, as shown in Figure 2.5-1, there is a great variety of perceived decision makers who decide on data retention policy. In the workshop discussions, it was suggested that in many cases these decisions are made by default rather than systematic process. Participants had only moderate awareness of the National Archives and Records Administration (NARA) role in records retention and whether data were seen as part of the records management process. NARA does not seem to provide any guidance on the valuation of data.

Valuation of data for unknown but potential secondary use is difficult. Very little helpful guidance exists on balancing cost of preservation and providing access with potential benefit from possible secondary use. Policy should encourage a repeatable process for valuation of data. Guidance on how to evaluate science data for long-term preservation should be developed and provided to the federal science and science data management communities. This could be a valuable guidelines program for NIST or more basic research programs for NSF. Emerging techniques such as social networking should be explored for use in data valuation.

Data curation includes the determination how long a dataset should be maintained before re-assessing its continued value. The determining factors of data curation for scientific data are recognized as inconsistent. Sometimes they are reflected in a records schedule, and sometimes they are dictated in a SDM Policy. Unfortunately, specific direction for data curation is often not mentioned at all.

As part of considering retention practices, participants acknowledged the difficulty in valuing benefits to secondary users. This particular concern becomes more acute as government leaders encourage reuse of data through initiatives like Data.gov. Participants felt that policy should be crafted that would assist federal agencies in making decisions with regard to the level of investment made in data access tools,

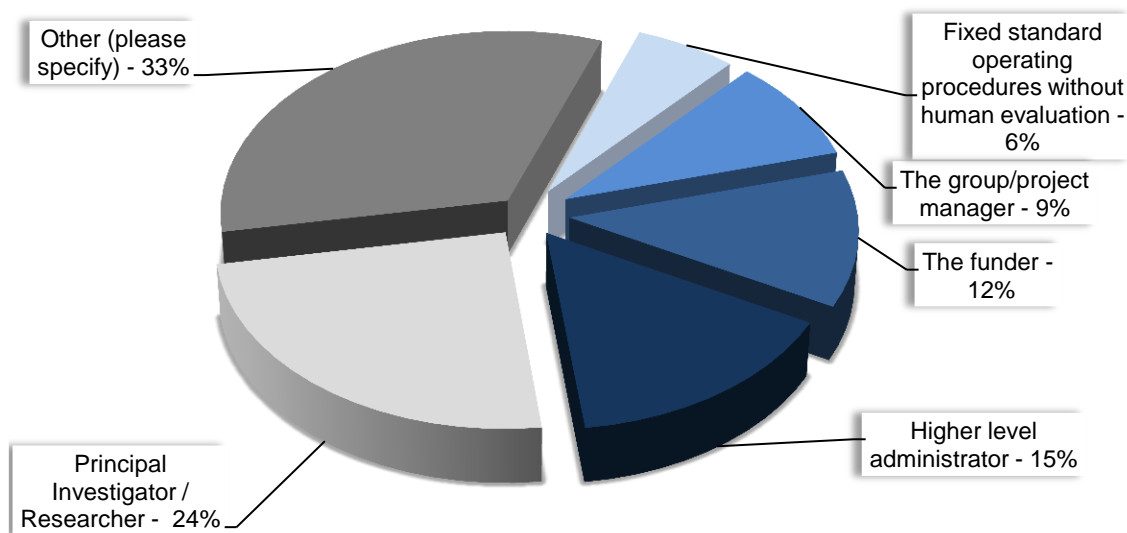
---

<sup>3</sup> *Harnessing Report*

<sup>4</sup> Two federal groups that share best practices in this community are the CENDI (the Federal STI Managers Group) and the Federal Library and Information Center Committee (FLICC).

metadata, and attributes of metadata that need to be provided such as data quality objectives. Participants also felt that periodic review and appraisal of digital assets should occur. Criteria should be considered to include whether data could be reproduced more cost effectively than archived, cost of preservation, and value to the broader scientific enterprise. Participants also suggested that datasets should undergo risk assessment to determine if keeping data, doing nothing specific to preserve or archive it, or deleting it given they were provided an optimum alternative. Discussion of how this review process would be implemented included consideration of the establishment of a review committee, which would include stakeholder representation. Such a process would be agency dependent. For further discussion, see Section 2.10 under Assignment of Responsibilities. Also mentioned was the potential use of social media to better understand potential secondary use and provide input to the data valuation process.

**In your agency, who decides if and how long data will be retained? (Q52, n=33)**



**Other**

- Combination of fixed standard operating procedure, funding entity, PI, and data stewards/archivists
- Established record schedule for the data
- Peer advisory panels
- Combination of NASA, data center managers, and users
- Researcher decides "if," archivist decides "how long"
- Appraisal archivists
- Records management schedules for scientific data

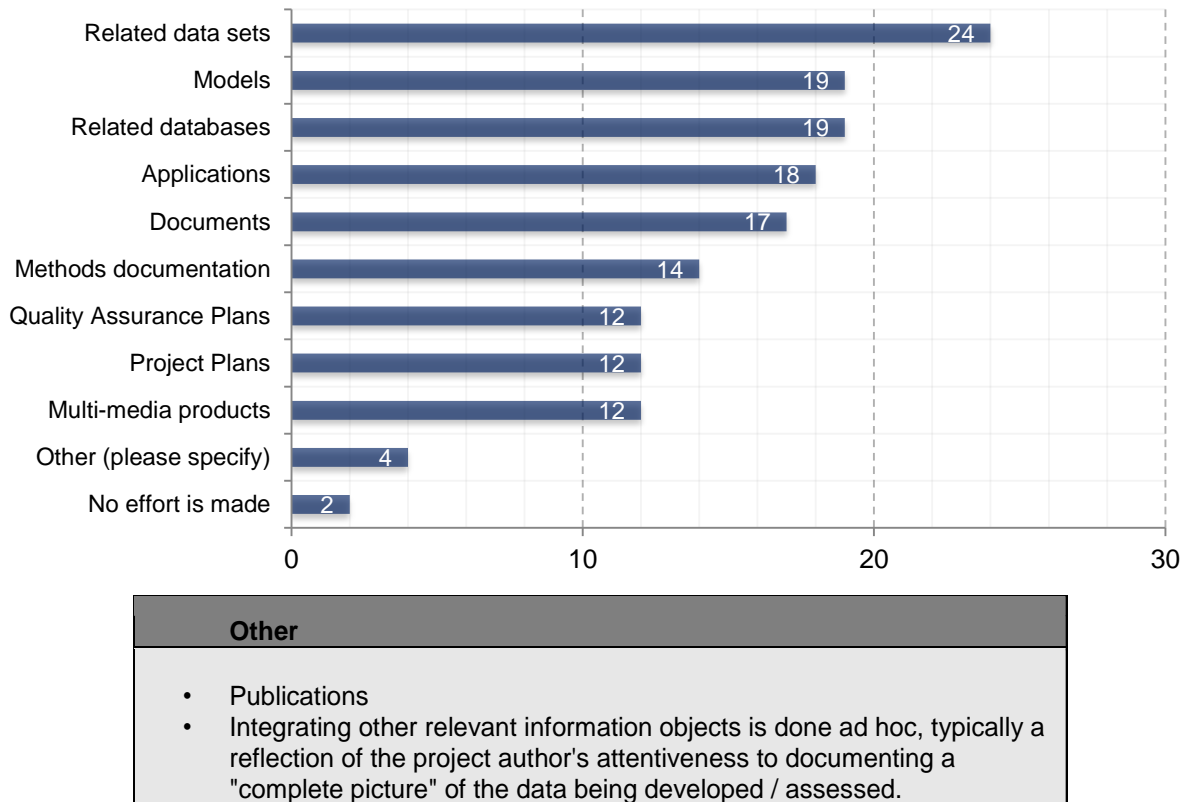
**Figure 2.5-1. Perceived decision makers on data retention policy.**

**2.6 Ensure that SDM Processes are Integrated with Knowledge Management Initiatives**

In workshop breakout sessions, it was explicitly stated that the concept of knowledge management was interpreted in different ways by different participants. The concept of integrating data into the knowledge management context was of moderate interest to participants, but enthusiasm was hampered because *knowledge management* is not clearly defined. However, there was strong agreement that there is

significant value in providing data in context. Figure 2.6-1 shows the types of linkages that participants felt were reasonably integrated in their agencies' data management processes.

**In a knowledge management context, in your community is there any effort to integrate data sets with any of the other relevant "information objects" listed below? Check all where there is integration of data sets. (Q58, n=31)**



**Figure 2.6-1. Information object linkages to data sets.**

Ideas that were considered included (1) a robust metadata system which uses managed vocabularies to control the variability of natural language, (2) technical approaches to link related digital objects, and (3) an interagency ontology. The ontology or domain model would be similar to that used in the federal enterprise architecture, and agencies could map their respective ontologies to the ontology. The capability to connect the SDM environment to the administrative environment, including budget and human resources, was also suggested.

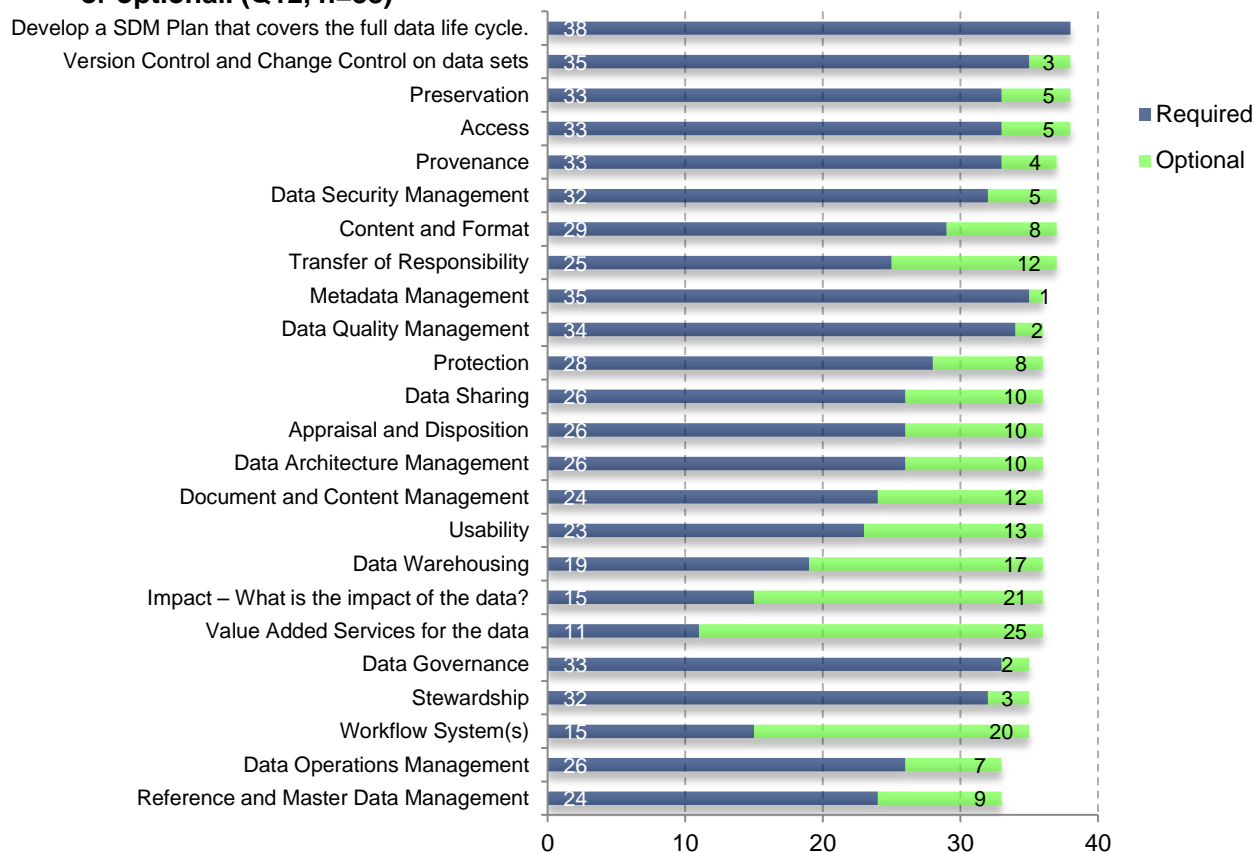
Policy should ensure that all science data are appropriately documented and managed within the context of the science project lifecycle, with all related digital objects linked to the datasets. Policy should also encourage presentation of high-value data in an application environment that provides as much additional context as practicable.

## 2.7 Identify Scientific Data with Metadata to Enable Needed Business Operations

Workshop participants recognized the critical role of metadata in enabling discovery, sharing, accessing, understanding, and using science data. The survey asked participants what elements should be in a data

management plan, and they ranked metadata management second only to covering the whole lifecycle as a required element. The full results for this survey question are shown in Figure 2.7-1.

**If you think about what elements should be in a SDM Plan for a specific program, grant, or research project, please check all that you would include as either necessary or optional. (Q12, n=38)**



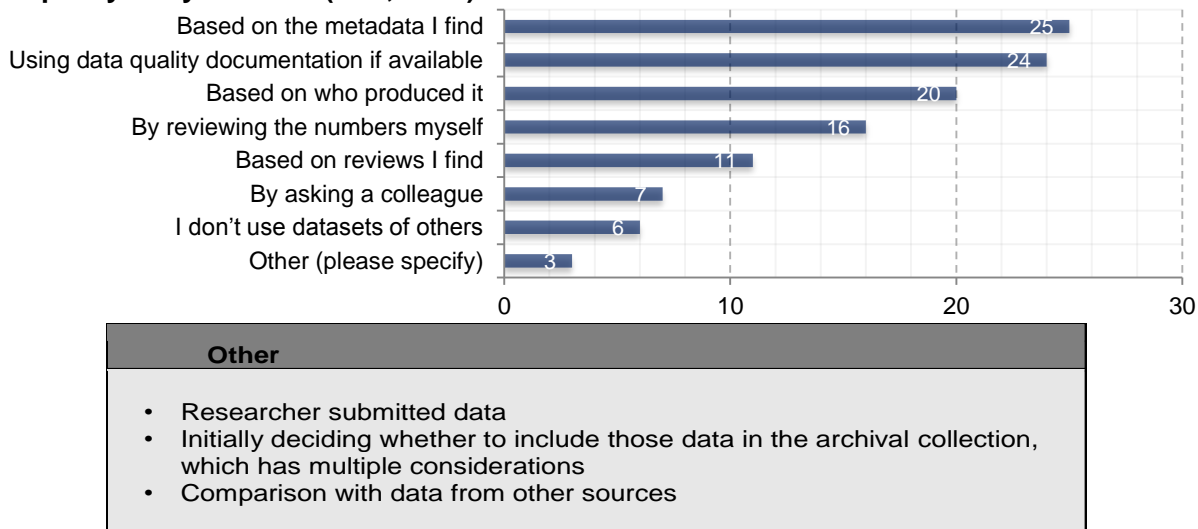
**Figure 2.7-1. Pre-workshop survey rankings of data management plan elements.**

In response to a question about how participants assessed whether data are of sufficient quality to use, the highest ranked source was “Based on metadata I find” (Figure 2.7-2 below).

The concepts of preservation and access are linked to metadata because preservation must include anything needed to enable access. These needs include metadata, tools that can access the metadata and data, and tools used to create, analyze, and model the data. Metadata should be linked directly with data, and technologies should be identified and implemented to enable linking and discovering data. Ontologies should be established or identified and implemented to enable and facilitate linking and discovery of data. Metadata should be developed and provided early in project and data lifecycles. Periodic review and appraisal of metadata should be conducted, and based on this process, metadata could change. It was also recognized that COPs should have the latitude to develop metadata structures with extensions and that persistent unique identifiers for digital objects are a critical component of data and metadata.

Questions on the survey addressed the issues of barriers and facilitators to secondary access to data. Of the survey participants, 67% stated that there are problematic barriers in scientific data access and exchange among agencies outside their COPs. Most notably, insufficient metadata to support use ranked as the most popular barrier (77%).

**In using data sets of others, how do you assess whether the data are of sufficient quality for your use? (Q24, n=36)**



**Figure 2.7-2. Participants’ perspectives on the sufficiency of data quality for use.**

It is difficult to anticipate secondary and tertiary uses and, by extension, long-term metadata requirements. Metadata needs may differ by user group, discipline, and domain. Metadata may need to be updated or added at the end of a project or at other points in the lifecycle. Fully documenting data and related information objects to support anticipated and unanticipated secondary use is currently expensive.

Based on the valuation of data, policy should encourage development of metadata that is sufficiently robust to support secondary use by applying appropriate community standards. Survey results indicated the need to implement standardization of metadata across information systems, as well as to provide data quality documentation (67%) to validate whether data are sufficient for agency use.

## 2.8 Manage Scientific Data for Appropriate Control

In the pre-workshop survey, approximately 60% of respondents indicated that their agencies placed controls on unclassified data, 16% indicated that their agencies did not, and 22% indicated that they were unsure. Furthermore, survey respondents provided a long list of classifications that they do use, including the following:

- Sensitive but unclassified
- Agency deliberative, confidential business information
- Contract sensitive
- Deprecation
- For official use only (FOUO/OUO)
- Proprietary
- International traffic in arms regulations (ITAR)
- Secret
- Agency only
- Competition sensitive
- Program use only
- Regional use only
- DoD Directive 5230.24

Workshop participants generally agreed that federally funded data ultimately belongs to the taxpayers, and as a result, policy should lead to the minimal level of control required to protect security, intellectual property rights, and privacy while maximizing benefit to the scientific enterprise as a whole. The concepts of embargo periods, time periods during which data are held by scientists to complete their research and publication efforts, and sunset dates when data are fully released, were all acknowledged as part of the solution to the challenge of managing data for appropriate control. Participants expressed concern that the term “sunset” may mean different things to different communities, with the suggestion that providing a glossary of SDM policy terms would ameliorate this impact. Finally, the group generally agreed that scientific data created by an agency should be made publicly available in a timely manner, but that there is a need to clarify what data to release and to clearly define who makes decisions on when to release controlled data. Clearly some data would always remain restricted with no release date, but this should be a conscious decision rather than protection of data by default.

The landscape of classification categories is difficult to navigate and seems to be applied in a non-standardized manner across agencies. Furthermore, there is a dearth of long-term repositories available to provide access to scientific data under a homogeneous classification system.

In response to the question of whether users in the operational environment encountered problems in accessing data they needed due to controls placed on the data by data producers/providers, over 52% said yes. In response to the question, “How easy is it to know about rights and restrictions on the use of data?” there was a mixed response, but 39% responded that it is difficult, and 17% responded that it is moderately difficult, for a combined, 56% responding that it is difficult or moderately difficult to know about rights and restrictions on data. When asked if there are needed sources of data currently unavailable or difficult to access and evaluate, over 65% said yes. Some of the reasons given for this difficulty were insufficient metadata and cultural roadblocks (i.e., data hoarding, treating data as proprietary, negotiations, and “red-tape”).

Policy should be used to clarify and balance access with control of data. In addition, policy should catalyze development of incentive structures that reward researchers for sharing data.

## **2.9 Maintain Version and Change Control on Datasets**

When science is conducted in a digital environment where science objects are linked electronically, data management becomes critical to both legal defensibility of policy decisions and scientific reproducibility of results. Time and money are saved when people are not duplicating effort unnecessarily. Participants generally agreed that maintaining version and change control on datasets and related objects is important. Operational users in particular commented on the importance of linking the correct versions of models to their corresponding input parameters and outputs. Operational users were also interested in ensuring support to the critical functions they perform by providing appropriate versions of the data and applications that they require.

Over 64% of those surveyed indicated that there are problems in their environments associated with accessing the appropriate version of data. It was strongly noted that lack of a persistent identifier approach for the federal sector makes linkage of digital objects difficult. Although linkage is occurring within some disciplines and COPs, best practices are being adopted slowly. Change control on datasets requires continual effort and may not be feasible to maintain once projects end.

It was clear from discussion and survey results that federal-wide SDM policy should encourage the use of a uniform approach to persistent digital identifiers that includes a means to support version control.

## 2.10 Assignment of Responsibilities

*The roles of agency offices and officials in implementing the agency digital data policy should be described to ensure clear lines of authority and accountability and to provide transparency for those working within and outside the agency on digital data matters. This should include provisions for a designated, cognizant senior science official serving as Science Data Officer to coordinate the digital data activities of the agency and to serve as representative to the Subcommittee on Digital Scientific Data. (NSTC, 2009)*

Participants affirmed the need for policy to clarify the roles and responsibilities of agency personnel. Some felt strongly that the position of chief data officer (CDO) was key to deriving full value from scientific data and data assets. Concerns were expressed regarding the real cost to the agency of implementing a CDO position, defining the relationship of the CDO to the CIO, and the relationship of SDM to enterprise architecture. It was felt that well-crafted policy could address these concerns and should affirm CDO authorities and responsibilities and those of related officials.

Regardless of whether this is a new, separate position or whether it is integrated with the duties of an existing C-level official, policy should be developed to encourage development of the functions of a CDO position and, to the extent possible, the clear, visible assignment of those functions. Policy should address target outcomes for data stewardship to which agencies should commit to achieve.

The development of CDO as a position may be resisted by some agencies. Resources for personnel required to provide full data curation and reference services may be difficult to obtain. Evolution of the research culture to more fully accept these responsibilities is expected to be slow.

Participants also delved deeper into the organizational structure in their discussions. They acknowledged the need for continuously funded data curation and stewardship functions and staff within the agencies to ensure continuity of SDM. They suggested that data stewardship should be a responsibility at all levels of an organization, but that a policy does not necessarily ensure a cultural change in behavior. They recognized the need to incentivize and encourage data management planning at the appropriate levels within the agencies, to develop roles and responsibilities to provide reference services for scientific data, and to improve interactions among researchers, data managers, and data scientists. Particular concern was expressed regarding responsibilities for long-term management of data after project close or during organizational changes, such as when personnel with key understanding of data move to other assignments.

In addition to the role of a CDO, there was a breakout group suggestion that a review committee should be established to help make determinations about data retention and curation. The need for such a structure is agency-dependent, but it should be given consideration. The issue of broader stakeholder

### Addressing New Roles and Responsibilities: Environmental Data Stewardship at NOAA

The National Oceanic and Atmospheric Association (NOAA) has become a leader in environmental data management by placing the improvement of data stewardship among the organization's top priorities. NOAA has strengthened its policies and directives based on a framework that conceives of data management as an end-to-end process and recognizes data stewardship as the key activity that overarches data observation, integration, and dissemination; archiving; access; and use.

NOAA officials have identified a growing need for expertise in guiding their high volume of data – approximately 4-5 petabytes per year – through the environmental data management process. Toward this end, the agency established the Environmental Data Management Committee in fall 2009 to provide leadership and coordinate data management strategy, policy, guidance, and implementation across NOAA. For the future, efforts to educate a data management workforce have been emphasized, including partnering with Earth Science Information Partners (ESIP) to develop a one- or two-day course and practicum in scientific data management for graduate students and junior scientists.

*Sources: Workshop Presentation on NOAA Data Stewardship, Donald Collins, 6/29/2010; National Research Council, Environmental Data Management at NOAA, 2007.*

involvement should also be given consideration either in policy or in data management planning infrastructure.

### **2.10.1 Communities of Practice (COPs)**

As part of the cultural change at all organizational levels, there needs to be recognition that COPs with shared goals and objectives are key organizational structures that often span agencies. Policies must address SDM governance and responsibility for data within these units. The workshop found that policy should be leveraged to formalize development and recognition of COPs. In addition to enabling good governance and data stewardship, formalizing COPs was suggested as the most promising approach to allocating the cost of managing data across a larger group of consumers with interest in the data, potentially relieving individual projects of the full cost of managing data for secondary and primary use.

However, COPs that need to address SDM issues are dynamic. Many disciplinary COPs have “informaticists,” or computational scientists, as members. Many information COPs have significant interaction with COPs in the scientific disciplines. These intersections are fostering an evolution of the roles of data scientists and data managers.

### **2.10.2 Reward Structure**

The *Harnessing Report* emphasized that, as data become more complex and SDM becomes increasingly significant, new roles evolve that are encumbered with new responsibilities to be filled to ensure successful SDM. The *Harnessing Report's* Appendix C lays out roles and relationships by sector, professional orientation, and type of organization. Workshop deliberations supported the need for filling the types of roles described in the *Report*, but they also raised the problem of how these roles can be filled in era of declining resources.

It is important to understand the view of responsibilities through the lens of both the producer and the consumer of data. To effectively treat data as a national asset, producers of data need to expend resources to make it broadly sharable. Consumers of data benefit when extensive effort has been made by producers to document and make data accessible in a manner that provides sufficient context for secondary use. Participants in the workshop recognized that there is no current incentive structure to equitably distribute the cost of making data sharable among producers and consumers. Pre-workshop survey results indicated that just under half (47%) of respondents feel that there are currently no incentives to promote data sharing among agencies. Workshop participants discussed treating data publication of data in a way parallel to publication of articles in scientific or scholarly literature. This was offered as a model to be developed for data sets. Providing easy, standard means of citing datasets will ensure credits to the individuals and organizations responsible for the data and will encourage data sharing. These cultural changes to the incentive structure should be encouraged not only in the federal sphere, but also in academia.

Along with an individual award structure to ensure that data are discoverable, cost should be allocated to ensure the capacity to provide long-term preservation. Institutional repositories, data management by COPs, and federal-wide efforts such as Data.gov need to demonstrate benefits to those who provide documented data to them in order to be sustainable.



## 2.11 Statement of Intentions and Mechanisms for Cooperation, Coordination, and Partnerships

*The agency digital data policy should describe the agency's intentions and mechanisms for cooperation, coordination, and partnerships across sectors. Such sectors can include government at the national, state, or local levels, as well as industry, academia, education, non-profits, and international entities. (NSTC, 2009)*

Although this was a federal workshop, participants recognized that “preservation of digital scientific data is both a government and private sector responsibility and benefits society as a whole.” Science challenges do not defer to agency, public, private, or national boundaries. Science lives in a global context, and it increasingly requires cross-boundary cooperation. This is confirmed by the COPs that work within such a context.

There is a broad array of interest groups working on different aspects of SDM. Strategies must be developed and implemented that leverage public and private resources in a more coordinated and methodical manner that will expedite progress in meeting SDM challenges. Further, as agencies craft their own policies, interagency policy must serve to catalyze cooperation and coordination while remaining flexible enough to not inhibit innovation at the project level.

Government agencies' data policies support or impede recognition and formal support of COPs and effective data sharing and use. Unless required by law or national security interests, policy should support formalization and governance of COPs and data sharing in an inter-sector, inter-organizational context.

## 2.12 Provisions for Updating and Revisions

*The breadth and dynamic nature of change in the technological environment was recognized, as was the fact that SDM policies and plans must be crafted to accommodate it. (NSTC, 2009)*

“Dynamic strategies are required.”<sup>5</sup> Workshop participants accepted that one goal of federal SDM policy should be to leverage new, emerging technologies in a cost-effective way. To achieve this goal, the following elements are necessary:

- Flexibility in policies and data management planning should be balanced with requirements for interoperability, access, and preservation. Policy, procedural, and technical solutions can help to achieve the balance, but they need to be put into a framework and addressed at an appropriate level of detail.
- SDM approaches must be harmonized with changes to approaches in science. Large repositories of data provide opportunities for data mining and “mash-up” analysis that previously did not exist. Enabling these activities requires that data managers pay additional attention to documentation of data quality and the experimental methods which produced the data.

Agency digital data policy must be a living document if it is to remain relevant and effective in a dynamic landscape. The policy should describe mechanisms to be used for updating and revising the document to ensure that it is responsive to change and opportunity.

Participants agreed that policy should be regularly reviewed and updated, and this supports the idea of having a responsible data authority with expertise in the state of practice for data policy and data management.

---

<sup>5</sup> Guiding principle from the *Harnessing Report*.

### 2.13 Data Management Policy Outline

In addition to the framework discussion, throughout the pre-workshop survey, data were gathered regarding the structure for a policy. The results of what elements should be in an agency-level data management policy and whether they should be required or optional are presented in Figure 2.13-1. In addition, the response to survey question 11 suggested that 17 additional elements could be added. They are given in Figure 2.13-2.

**If you think about what elements should be in an Agency-level SDM Policy, please check all that you would include as either necessary or optional. Then add any elements not on the list. (Q10, n=36)**

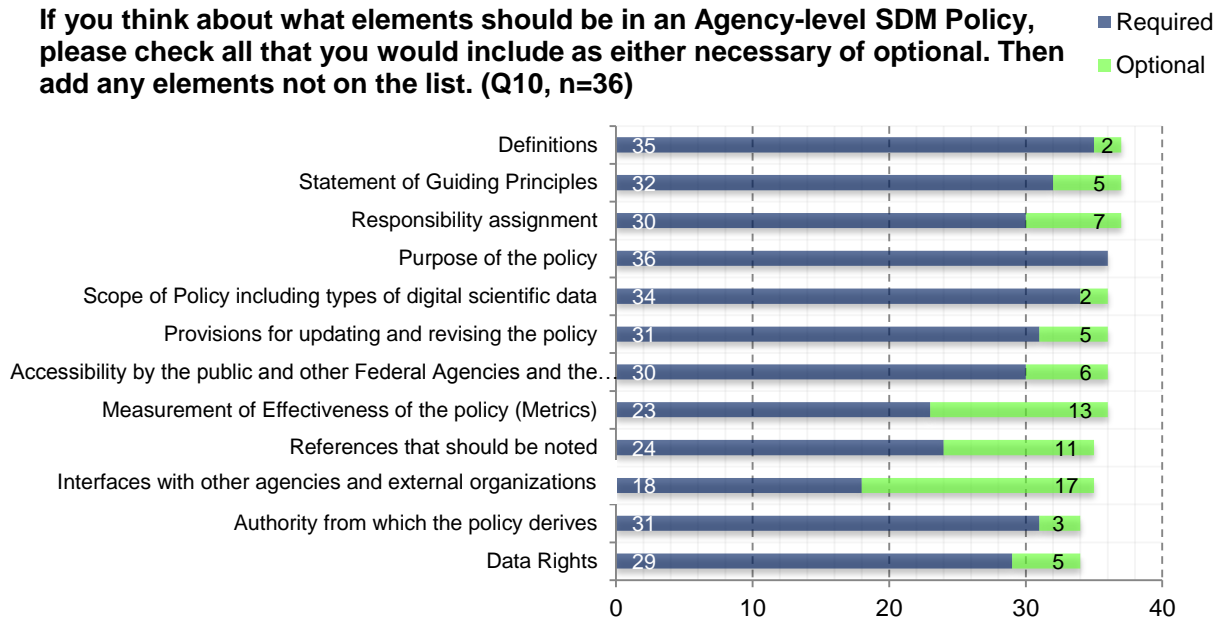


Figure 2.13-1. The popularity of required and optional policy elements requested for policy structure.

Required	Optional
<ul style="list-style-type: none"> <li>• Guidelines on preserving the provenance of data</li> <li>• Provisions of lifecycle cost and level of services for data management for projects</li> <li>• Data security and catastrophe planning</li> <li>• Confidentiality, integrity, and access (general and for people with disabilities)</li> <li>• High-level adoption of a data lifecycle management framework to manage data assets</li> <li>• Time frame for developing and updating the policy</li> <li>• Policy on using standards</li> <li>• Long-term preservation and archiving</li> <li>• Points of contact (individuals and agencies responsible)</li> </ul>	<ul style="list-style-type: none"> <li>• Association with derived scholarly communication auditing (data and security)</li> <li>• Data retention policy</li> <li>• Internal and external stakeholders</li> <li>• Security management</li> <li>• Linkages with other intra- and inter-agency efforts</li> <li>• Record of the resource impact that a policy can have</li> <li>• Policies as big cost drivers to projects</li> <li>• Points of contact (individuals and agencies responsible)</li> </ul>

Figure 2.13-2. Seventeen additional policy elements requested by survey participants.

The text box at the right contains a list of sequential headings for an SDM policy that was surveyed. Recognizing that each agency has a structure for its policies and directives, the data could be cross-walked to the agency's context.

Appendix C provides a sample data policy based on ideas from the workshop and practices and guidelines provided in the Workshop background material.

### Headings for an SDM Policy

- Authority from which the policy derives
- Purpose of the policy
- Statement of guiding principles
- Definitions
- Scope of policy, including types of digital scientific data
- Responsibility assignment
- Data rights
- Interfaces with other agencies and external organizations – statement of intentions and mechanisms for cooperation, coordination, and partnerships
- Accessibility by other federal agencies and the public
- Measurement of policy effectiveness (metrics)
- Provisions for updating and revising the policy
- References that should be noted

### 3.0 DATA MANAGEMENT PLANS

This section identifies findings for data management plans based on background material and discussions conducted at the workshop. Workshop participants were charged with identifying important issues and elements that should be considered in developing a data management plan.

The elements of a data management plan (shown in Figure 3.1-1) were provided to each breakout group. The elements and definitions were compiled from the *Harnessing Report*, the *EPA Survey*, and *DAMA's Functional Framework*.

Figure 3.1-1. Data management plan elements.	Goals and Principles	Roles and Responsibilities	Best Practices	Issues
<b>Description:</b> Brief, high-level description of the digital scientific data to be produced				
<b>Impact:</b> Discussion of possible impact of the data within the immediate field, in other fields, and any broader, societal impact. Indicate how the data management plan will maximize the value of the data.				
<b>Data Governance:</b> The exercise of authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets; high-level planning and control over data management. (DAMA)				
<b>Content and Format:</b> Statement of plans for data and metadata content and format, including description of documentation plans and rationale for selection of appropriate standards. Existing, accepted standards should be used where possible. Where standards are missing or inadequate, alternate strategies for enabling data reuse and repurposing should be described, and agencies should be alerted to needs for standards development or evolution.				
<b>Data Operations Management:</b> Planning, control, and support for structured data assets across the data lifecycle, from creation and acquisition through archival and purge. (DAMA)				
<b>Data Architecture Management:</b> Development and maintenance of enterprise data architecture within the context of all enterprise architecture, and its connection with the application system solutions and projects that implement enterprise architecture. (DAMA)				
<b>Version Control and Change Control on Datasets:</b> Control of scientific data to ensure the integrity of data and final products. Data within a project undergoes a continued development phase, from working data to mature, released, submitted, and archived data.				
<b>Metadata Management:</b> Planning, implementation, and control activities to enable easy access to high quality, integrated metadata. (DAMA)				
<b>Data Quality Management:</b> Planning, implementation, and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use. (DAMA)				
<b>Usability:</b> information about suitability of data for known or anticipated uses.				
<b>Access:</b> Description of plans for providing access to data, which should include (1) a description and rationale for any restrictions on who may access the data under what conditions and a timeline for providing access, and (2) a description of the resources and capabilities (equipment, connections, systems, expertise, etc.) needed to meet anticipated requests. These resources and capabilities should be appropriate for the projected usage, addressing special requirements such as those associated with streaming video or audio, movement of massive data sets, etc.				
<b>Data Security Management:</b> Planning, implementation, and control activities to ensure privacy and confidentiality and to prevent unauthorized and inappropriate data access, creation or change.				

Figure 3.1-1. Data management plan elements.	Goals and Principles	Roles and Responsibilities	Best Practices	Issues
<b>Protection:</b> Statement of plans, where appropriate and necessary, for protection of privacy, confidentiality, security, intellectual property and other rights.				
<b>Reference and Master Data Management:</b> Planning, implementation, and control activities to ensure consistency of contextual data values with a “golden version” of these data values. (DAMA)				
<b>Data Warehousing:</b> Planning, implementation, and control processes to provide decision support data and support knowledge workers engaged in reporting, query and analysis. (DAMA)				
<b>Document and Content Management:</b> Planning, implementation, and control activities to store, protect, and access data found within electronic files and physical records (including text, graphics, image, audio, and video). (DAMA)				
<b>Transfer of Responsibility:</b> Description of plans for changes in preservation and access responsibility. Where responsibility for continuing documentation, annotation, curation, access, and preservation (or its counterparts, de-accessioning or disposal) will move from one entity or institution to another during the anticipated data lifecycle, plans for managing the exchange and documentation of the necessary commitments and agreements should be provided.				
<b>Appraisal and Disposition Preservation:</b> Description of plans for preserving data in accessible form. Plans should include a timeline proposing how long the data are to be preserved, outline any changes in access anticipated during the preservation timeline, and document the resources and capabilities (e.g., equipment, connections, systems, expertise) needed to meet the preservation goals. If data will be preserved beyond the duration of direct project funding, a description of other funding sources or institutional commitments necessary to achieve the long-term preservation and access goals should be provided.				
<b>Stewardship:</b> Transfer of responsibility; description of plans for changes in preservation and access responsibility. If responsibility for continuing documentation, annotation, curation, access, and preservation (or its counterparts, de-accessioning or disposal) will move from one entity or institution to another during the anticipated data lifecycle, plans for managing the exchange and documentation of the necessary commitments and agreements should be provided.				
<b>Provenance:</b> Description of data history.				
<b>Value added services for the data:</b> Descriptions of transformations and other processing performed on data prior to or during its use.				
<b>Data sharing:</b> Data reuse and repurposing should be described, and agencies should be alerted to needs for standards development or evolution.				
<b>Workflow system(s):</b> Applications that automate the flow of work to resources involved in the activity.				

**Figure 3.1-1. Data management plan elemental matrix composed from three resources: *The Harnessing Report*, *EPA Survey*, and *DAMA’s Functional Framework*.**

### 3.1 General Considerations for Science Data Management Plans

The concept of SDM planning is relatively new in the federal research community, although technical and administrative data management approaches and standards provide some framework. For example, ANSI standard, ANSI-GEIA 859, *Data Management*, is a data management process standard that EPA ORD has adapted for its SDM policy. This standard is also extensively used in the DoD/NASA technical data management community. There is also a COP for data management: the Association for Configuration and Data Management. A COP to describe the conceptual framework for scientific data and to provide guidance on its implementation, particularly as it has important differences from other forms of data. A fully described set of SDM functions, or a model of the overall SDM process, could help address the community's uncertainty. For example, there is a relationship between the size, complexity, number of data sets, access requirements, processing requirements and preservation needs of data generated by a project and the scope of the required SDM Plan. Within science data, there are also significant differences in the way that high-volume data streams (such as those from satellites) are managed versus small-volume data collections managed by single investigators. There is little current guidance to assist scientists and data managers in determining SDM plan scope. There is also a concomitant need for training in SDM planning, which could be augmented by examples of effective SDM plans from particular domains or COPs. Guidance is needed on how the SDM planning function will interact with other agency-wide elements such as project management, quality assurance, and enterprise architecture. Existing governance processes (such as those included in the CIO function and enterprise architecture) need to be reconciled with SDM governance.

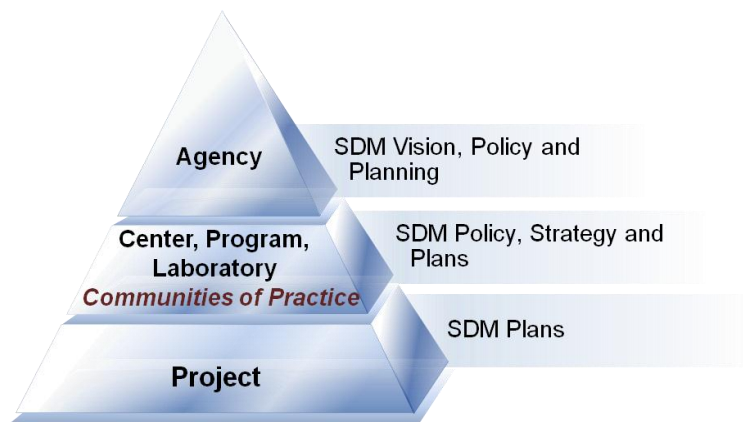
**Examples of Data Management Plan Templates**

The links below provide model plans developed for specific programs or communities of practice. These and other plans can be found in the SDM bibliography ([link](#)).

- Table of contents - Appendix D
- Digital Curation Centre template ([link](#))

There is wide-ranging diversity in the ways that projects are configured to meet science challenges. For example, the data management approach used to support mapping of the human genome may be quite different than that needed by a principle investigator making seasonal transects of the Chesapeake Bay. These differences dictate that one size or type of data management plan will not fit all applications. A sample SDM plan table of contents is provided in Appendix D and should be considered as starting point to be tailored for a diverse set of SDM implementation scenarios.

Agency-level SDM policy should encourage data management planning at all levels of the organization, to include project, program, agency, and COP levels. The nature of data management plans will range from being tactical plans at the project level to strategic plans at the agency level. Planning at all levels should contribute to efficiency in data management. As with policy development, as planning becomes more strategic, it should evolve to be more outcome oriented rather than process based. Figure 3.1-2 depicts the tiered nature of federal data management planning functions. At the agency level there is a vision and policy for SDM, and in some cases there is a strategic plan for its implementation agency-wide. Levels below the agency tiers configure their

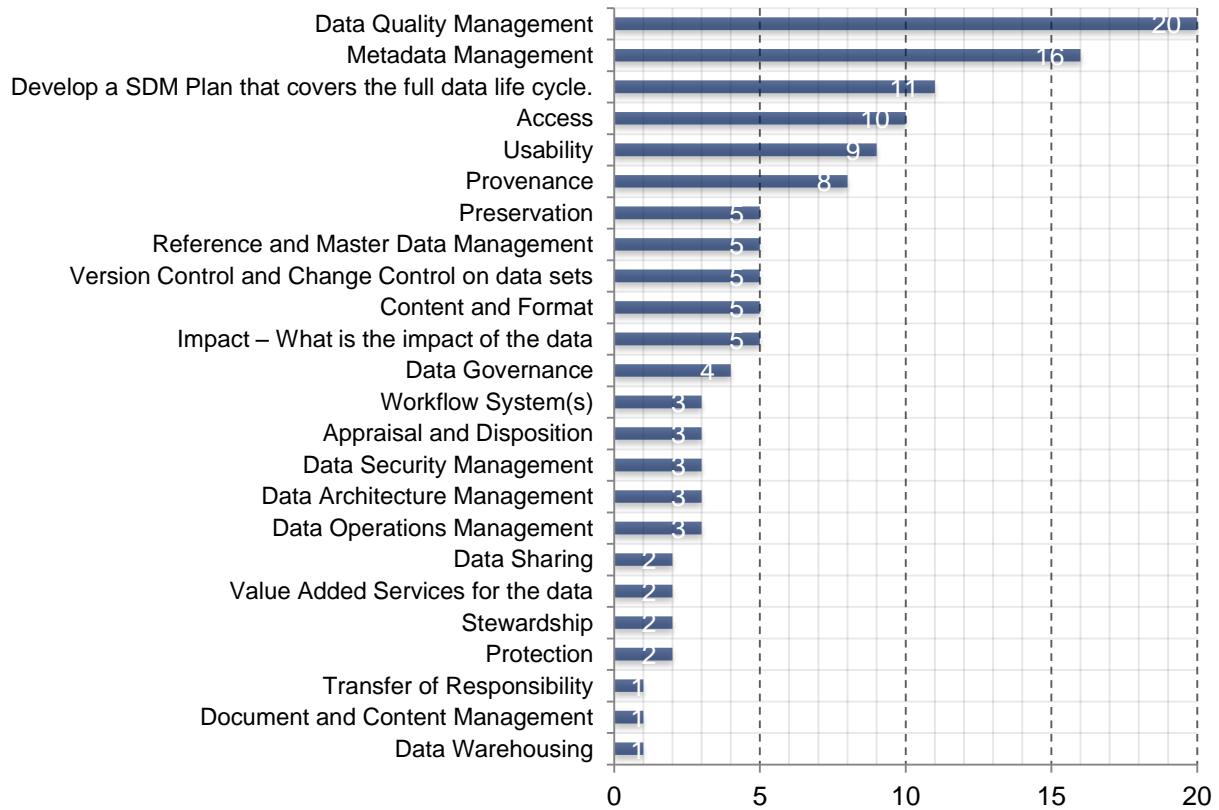


**Figure 3.1-2. A hierarchical structure of data management planning functions.**

strategies to contribute to the goals and objectives articulated at the agency level. The foundation on which these upper tiers depend is the project/data-set level, where data are initially produced and consumed. Likewise, almost all projects benefit from the capacities provided by upper tiers, including governance, stewardship, communication, and publication. Policy that provides the guiding principles for governance of SDM and incentives leading to appropriate stewardship of critical agency assets need to be inherited by projects from the agency level and augmented at the programmatic or COP level. In addition, it would be inefficient for each project to create and maintain architecture for communication, outreach and publication of its results. Ultimately, most projects leverage policies, plan components, and capabilities from all tiers in order to accomplish their data management planning objectives and to also address the full data lifecycle.

Participants were asked to select three elements of a data management plan that they find most helpful in evaluating data for use in the development of policy. Their responses are presented in Figure 3.1-3 below. Knowing how the data were managed to ensure quality ranked as the highest element, followed closely by metadata management.

**What 3 elements of a SDM Plan would be most helpful to you in evaluating data for use in making and defending policy recommendations? (Q29, n=36)**



**Figure 3.1-3. Prioritized data management plan elements most useful in defending policy recommendations as denoted by survey participants.**

Participants were aware that data management plans would vary based on the situation within which they were developed and implemented. Variables discussed included size and complexity of projects, whether projects were intramural or extramural, initial assessment the data’s value long term, size of the data sets, and availability of repositories or archives. Regardless of the situation, participants embraced the concept that the scientific enterprise would benefit from data management plans implemented as living electronic records rather than traditional documents. This concept was affirmed by 92% of survey participants.

In general, interpretation of what constitutes a living SDM plan varied from an electronic document continuously revised throughout the lifecycle to a set of related digital objects—such as project plans, data sets, models, publications, and the metadata describing them—continuously updated throughout the data lifecycle. The latter approach has the potential to enable secondary users to explore the full pedigree of data. However, to be effective, a persistent digital object linking infrastructure would be critical. It is important to plan what metadata will be generated, what format standards they follow, and other factors during the project planning phase to allow potential secondary users to anticipate data that will be available to them in the future. Further exploration is needed to determine implementation approaches to data management planning and guidelines that would be applicable throughout the federal sphere.

In order to achieve a comprehensive approach to data management planning that is integrated into the project lifecycle, a future state of compatible repository management approaches and standards is needed throughout the federal science agencies. Participants determined that an initial productive step would be to make a repository of best practice data management plans. Many plans' citations are contained in the bibliography created for the workshop ([link](#)), but a targeted list of best practices or exemplary plans needs to be culled out.

### 3.2 Observations on Specific Data Management Plan Elements

Figure 2.7-1 lists elements from these sources and shows attendee survey results indicating which elements should be required and which should be optional. Almost all the data elements listed were considered useful, but many evoked sufficient support from respondents to be considered for mandatory inclusion as elements in a data management plan.

Respondents also indicated in Figure 3.2-1 what they thought were the key barriers to secondary use of data. Attendees' choices were consistent with the previous responses and emphasized the critical role of metadata development and management, as well as the linkage of related project digital objects in providing a comprehensive data pedigree.

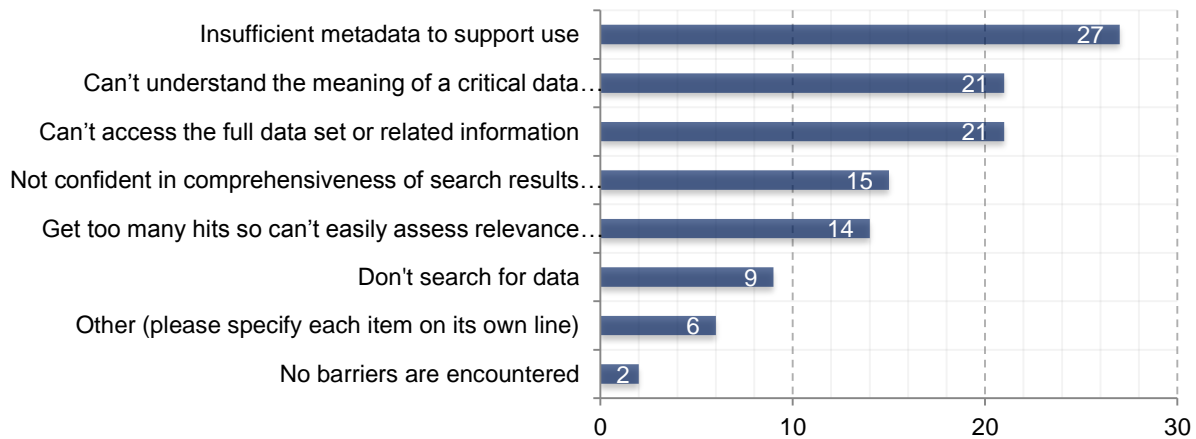
All survey participants indicated that projects should be required to develop a data management plan that covers the full data lifecycle. Most of the SDM plan could be drafted by data management personnel and then completed in cooperation with the principal investigator and team. SDM plans need to be adaptable to accommodate project changes, and they should be stored in a repository accessible to project personnel for reference and update. They should also be available to other agency staff for review. Upon approval, SDM plans should be stored in a location accessible not only to the project members, but to other project leaders and/or principal investigators looking for examples of SDM techniques.

The importance of metadata dictates that metadata standards be identified in the plan. If metadata standards are not used, the plan should have a description of the content and format of the data; this should cover as much of the data lifecycle as is known when the plan is written or revised. It should describe how metadata will enable discovery of the data set after the project closes. The latter aspect may not be known until later in the project lifecycle, so this section of an SDM plan may be a placeholder until that time.

The following sections address workshop observations on specific elements of the SDM plan. The percentage associated with each element represents the number of survey participants who responded affirmatively that the element should be considered as a required element in the SDM Plan.



**What barriers do you encounter when searching for or re-using “found” data generated by others? Check all that apply. (Q23, n=35)**



- Other**
- Linking of publication with data source
  - Federated search across information systems
  - Need to derive metadata from existing metadata
  - Getting permission to access this data
  - Find multiple versions with uncertain lineage
  - Incorrect or misleading metadata
  - Dead links in metadata
  - Standardization of metadata across information systems
  - Insufficient information on QA/QC procedure details
  - All items listed above are cited as reasons for not finding data in our systems.

**Figure 3.2-1. Most identified barriers for the secondary use of scientific data.**

**3.2.1 Description and Impact (30%)**

It was universally accepted that data management plans should describe the data produced by a project. Over 97% of responses ranked “definitions of data elements” as the most desired level of documentation to support the use of “found” data. This **description** would likely consist of general information about the data contained in a textual abstract with detailed information such as formats, lengths, and definitions of individual data elements contained in sections of the plan devoted to metadata and content and format descriptions. Source material used in preparation for the conference indicated that an assessment of the data’s **impact** was an important part of descriptive information. This did not seem to be supported by pre-conference survey results, with only 30% of respondents indicating that impact should be a required element of a data management plan. The probable cause of the practitioners’ ambivalence in this area is the difficulty in anticipating secondary use of data.

**3.2.2 Data Governance (84%) and Stewardship (81%)**

Institutionalizing the function of SDM planning in the federal sphere will require agencies to embrace the concepts of data governance and stewardship. In the pre-conference survey, both of these sections garnered significant support as mandatory elements in data management plans. Discussions in the breakout sessions recognized that the breadth of these elements requires various aspects of them to be addressed at the agency- and COP levels, as well as in project data management plans.

The concepts of governance and stewardship inherently include the allocation of responsibilities. Although different models were discussed, they all typically included a senior scientific data management official or chief data officer, some type of decision-making body or Scientific Data Management Council (SDMC), principle investigators, and scientific data stewards. Effective governance and stewardship result from the interaction of these groups and individuals, empowered by sound policy. The assignment of responsibilities discussed in a policy (See Section 2.10) should set the stage for agency data governance and stewardship.

**Governance** functions that generally should be addressed at the agency- and COP level include setting SDM policy, determining the scope and timing of the agency SDM implementation, making decisions and providing guidance on agency SDM practices, and determining whether to implement SDM only for new projects or to apply the approach to legacy data. Because of the scope, complexity, and critical nature of data governance in accomplishing agency missions, the concept of agency chief data officers and some form of agency SDM governance board was endorsed.

Governance and **stewardship** addressed in the data management plan involve leverage of personnel assets to perform data management functions to accomplish plan objectives. Approaches discussed in the breakout sessions included training existing science personnel to perform data management functions, including data management personnel from the outset of the project, and leveraging assets and capabilities in related programs like quality assurance (QA) and enterprise architecture (EA).

An example of this type of inter-program activity is the EPA's ORD, which requires projects to have a quality assurance project plan that addresses some of the elements of an SDM plan. When mandated program plans already include QA and EA, they can be incorporated by reference. The benefits need to be evaluated in light of the resources required to coordinate the efforts.

### **3.2.3 Data Sharing (68%), Access (84%), Data Security Management (81%), and Version Control (92%)**

Across the federally funded research community, many researchers regard data generated from their research as their exclusive property. This resistance to **data sharing** is especially prevalent within the agencies that do not provide repositories for research data and whose extramural agreements contain no discussion of data rights. However, there are many excellent examples of commitments to data sharing, such as that described in the IceBridge example in the call out box in this section. Pre-conference survey results indicate that 78% of respondents ranked resource concerns as the first or second impediment to managing data as an enterprise asset. This is a major barrier for data sharing. A similar percentage identified cultural concerns as a barrier. The practice of ensuring that the research team has exclusive access to the research data for primary publication is very common. A suggestion from the breakout sessions was to establish initiatives to help the federal research community become conversant with a more appropriate concept of "data rights" that provides adequate protection of intellectual property while providing access to the research community at large and the ultimate owner of federal data, the taxpayer.

A significant majority of survey participants indicated that access should be a required element in data management plans. Workshop participants agreed that scientific data access control and **data security** concerns stem from privacy, confidentiality, intellectual property, and other requirements. Sixty-eight percent (68%) of survey responses indicate that there are problematic barriers to scientific data access and exchange among agencies and external COPs. Although the majority of results from research funded by federal agencies are public data, some data need protection, and the integrity of all scientific data must be assured. To specify access control processes, an SDM plan must have complete information regarding different parties' rights to data. This helps to identify which data require protection from unauthorized access throughout the data lifecycle. SDM planning also requires review of all agreements, such as grants, cooperative research and development agreements (CRADAs), and interagency memoranda of agreement or understanding (MOAs or MOUs), that impact the project. In addition, other data protection triggers,

such as human subject research, personally identifiable information (PII), and confidential submissions from businesses must be considered.

Other complications discussed include changes of requirements for who can access data throughout the data lifecycle. These changes require SDM planners to be aware of who will access the data, how the data will be accessed, where it will be stored, the size of the data transferred and stored, and how frequently data will be accessed or downloaded.

Workshop participants were interested in the use of embargo and sunset periods on scientific data to simplify **preservation** planning (see section 3.4.5). Embargoing research data for later release (an event-based example is when articles containing the data are published) can be an efficient way to ensure that data are discoverable by other users while providing protection for intellectual property. Data may also be embargoed from release to certain entities. For example, certain technology information may not be released to non-US citizens. Participants acknowledged that establishing time- or schedule-based sunset periods on research data is an efficient alternative approach to setting dates for disposition or retention review. Data management plans need to account for the policies, guidelines, and mechanisms required to communicate about the availability of data, information, and tools. Data management plans should also provide this information with appropriate contextual data.

**Version control** is a significant consideration relating to security management and access. For example, sensitive data may be dismissed from a data set, allowing its classification to shift from restricted to public release data. Such subtleties can bear significantly on the overall classification of data, thus altering accessibility and potentially changing security management measures. The pre-conference survey asked participants whether they are aware of procedures for change control of data during planning and management of science projects. While 34% of respondents said that they are not sure, and 18% said no. In combination, over half indicated that they were not sure or that procedures do not exist. Only 18% said procedures do exist. Clearly, improvements in training or more and better procedures are needed in this area. This should be prominently highlighted based on the strong support (92% of respondents) for the requirement of version control.

In the operational environment, 40% indicated that there is no process to capture new data sets if they are modified or improved or if newly derived products are created. Thirty-one percent (31%) were not sure,

### NASA's IceBridge: An Airborne Mission for Earth's Polar Ice

Operation IceBridge is a NASA airborne mission to observe changes in Earth's rapidly changing polar land ice and sea ice. The mission is now paralleled by a campaign to bring data to researchers as quickly as possible and to accelerate the analysis of those changes and how they may affect people and climate systems.

"Anyone can access the wealth of IceBridge data online, and do so free of charge and without a formal request," said Michael Studinger, IceBridge project scientist at NASA's Goddard Space Flight Center in Greenbelt, Md. "It's critical for data to be free and accessible so scientists can conduct timely studies of ice dynamics and a changing climate."

To date, NSIDC has published 12 datasets from the IceBridge Greenland and Antarctica campaigns in 2009. These datasets spanned 10 instruments, including LIDARs, radars, sounders, gravimeters, mappers, and cameras, as well as atmospheric measurements and aircraft positioning data.

"It's exciting to have such a diversity of data, preserving it for the future and making it available in ways that will encourage new discoveries," said Marilyn Kaminski, NSIDC's project manager for IceBridge. "There's so much potential that can be tapped."

NASA flew its 2010 IceBridge Greenland campaign from March through May; data will be available at NSIDC in fall 2010. NSIDC will publish data from subsequent campaigns within six to eight weeks of receipt from the data providers. This rapid turnaround will enable researchers to use these important data to monitor receding glaciers, the melting Greenland ice sheet, crumbling ice shelves on the Antarctic Peninsula, and the thinning of old, thick Arctic sea ice that has been the mainstay of the sea ice cover.

Source: [http://www.nasa.gov/mission\\_pages/icebridge/index.html](http://www.nasa.gov/mission_pages/icebridge/index.html)

for a combined 71% negative response. These responses may indicate that the community has experienced a significant loss of effort and useful data products. This issue should be explored further.

### **3.2.4 Metadata Management (97%), Content and Format (74%), Document and Content Management (63%)**

Section 2.7 of this report discusses metadata and its importance in policy considerations. It also addresses assessment of quality. Survey participants almost universally acknowledged **metadata management** as a required element of an SDM plan, despite the producer's perspective that resources expended to document data for secondary use imposes an additional burden on projects. With notable exceptions (e.g., the Federal Geographic Data Committee (FGDC)), there is little guidance on producing metadata sufficiently robust to support secondary use. This impedes discovery and use of data, and it causes a reliance on peer networks and indirect approaches to data discovery through publications, reports, or other products that result from the original research project.

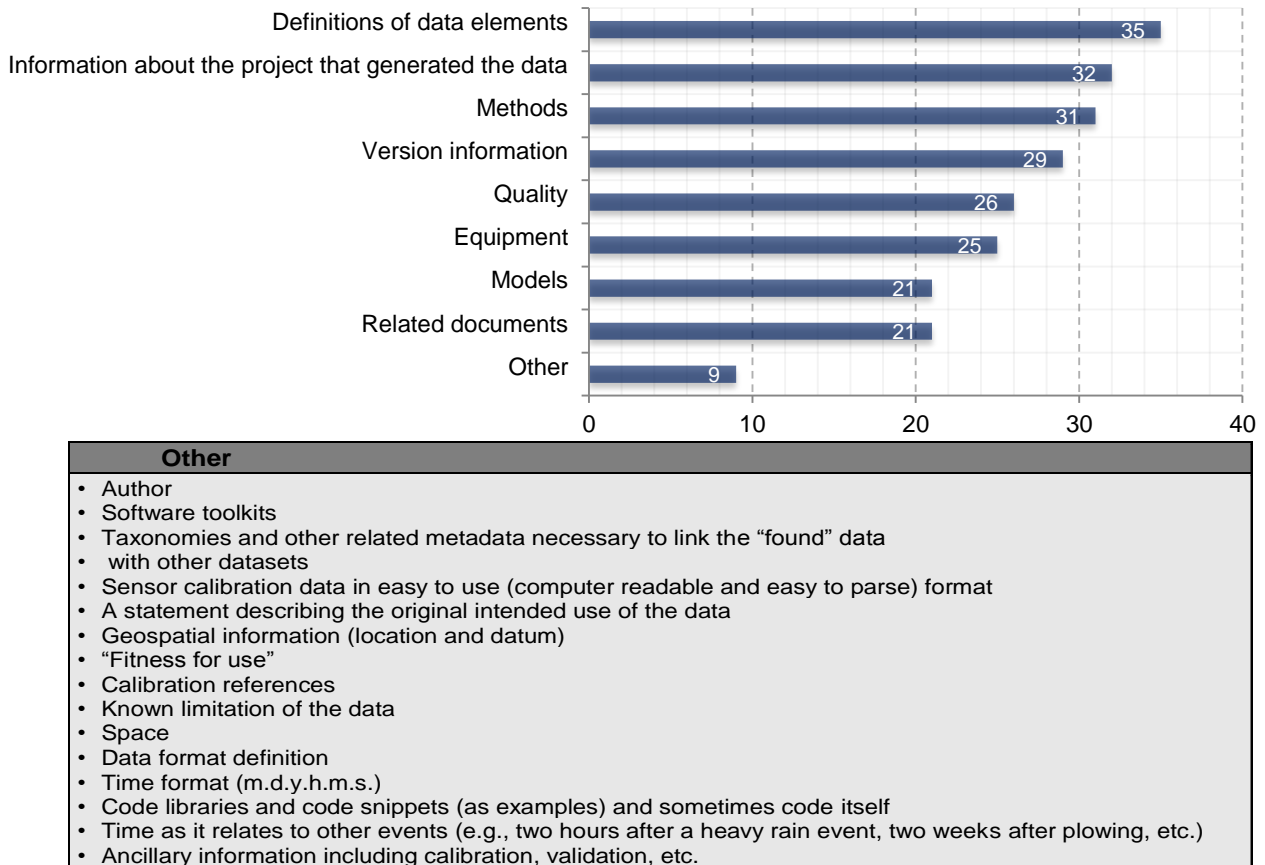
Based on the survey, it was evident that there are significant barriers when searching for "found" data generated by others:

- 77% responded that there is insufficient metadata to support use when searching for "found" data generated by others.
- 60% of responders indicated that they cannot access the full data set or related information.
- 60% also said that they cannot understand the meaning of a critical data element(s) (i.e., variables or no defined meaning).

This led to the conclusion that the research community needs guidance on metadata required for others to identify their data sets and make those data discoverable, understandable, and usable. One suggested solution to the metadata challenge was to allow archivists and librarians to serve as metadata managers.

Figure 3.2.4-1 provides information on the types of metadata needed to allow for use of data generated by others. The question treated scientific data as monolithic, so it may not have accurately addressed the variation in metadata preferences based on specific types of scientific data.

**In your research community what level of documentation do you typically need to support use of “found” or other people’s produced data? Check all that apply.**



**Figure 3.2.4-1. The level of documentation needed to validate the quality of “found” data.**

Use of standards, when they exist, helps to ensure that data sets are managed as enterprise assets. Standards that apply to the project need to be documented. Many types of standards may be relevant, but at a minimum, the SDM plan should document the standards that impact data during their lifecycle. This would include those for vocabularies, taxonomies, quality, and methods. There may be applicable standards for scientific data capture; preparation such as coding, record, or database format processing; analysis; data transfer; or storage. When used, standards should be cited with complete information. An example of one such standard is ISO 10390:2005, *Soil quality - Determination of pH*. Planning efforts should also take into consideration that standards change over time. Use of standards referenced in the metadata describing projects, data sets, and other related digital objects can expedite metadata creation and can result in data with more interoperability.

**Content and format** includes addressing the actual structure of the data and metadata. Fewer survey participants viewed this as mandatory when compared to metadata management. This may reflect the lack of adoptable metadata standards in many disciplines and COP. The definition used for “content and format” may suggest a level of structure for the data and metadata at the initial planning phase that would be difficult to meet without the use of standard formats.

Because breakout sessions were composed of researchers, operational users, science managers, and policy analysts, the metadata requirements that were needed to satisfy the requirements of these groups were expected to differ. For example, the rigor of research might require awareness of data coverage, precision, methods, assessment of usability, access restrictions, security required once downloaded, and full provenance of the data. In contrast, operational users with thoroughly vetted data in an operational environment may have more interest in those elements that enable rapid discovery and access.

**Document and content management**, as defined for the workshop, referred to data imbedded in other digital objects, including publications, video, audio, and other multimedia products. In order to effectively address management of these objects within the project and data lifecycles, metadata standards need to be developed to describe the objects, and they need to be labeled with a persistent digital identifier. The NIST presentation to the plenary session demonstrated the value for one COP as being able to mine publications for data imbedded therein. The presenter described a set of applications developed and used by the Thermophysical Properties Division. These applications ingest publications, extract their thermophysical property data, and use the resulting database to provide feedback to investigators on inconsistencies in their data. Pre-survey results indicated that a little more than half of survey participants thought that document and content management should be a required element of a data management plan. The interpretation of this result could be that some participants infrequently use this sort of digital object in research, or that the lack of metadata standards and digital object identifiers makes it unrealistic at this stage of development to effectively plan for the use of this type of data.

**Linking Data to Projects in Repositories and Archives: Complex Astrophysical Data Systems**

The ability to connect scientific results with the data and processes used to produce them represents a crucial part of research infrastructure and becomes more difficult as amounts of digital data rapidly expand. The Smithsonian Astrophysical Observatory (SAO), under a NASA grant, has created an innovative solution: the Smithsonian / NASA Astrophysics Data System (ADS) is a digital library portal maintaining a bibliographic database of millions of records for scholarly literature, citation information, and full-text historical astronomy publications. For each article in its database, the ADS provides access to its metadata, citations, readership statistics, and external resources, including electronic data catalogs and archives.

One project successfully linked with the ADS is the Chandra X-Ray Center (also a NASA-funded SAO project). The center manages the observations of the Chandra X-ray observatory as it investigates phenomena in space. The ADS provides links to the official archive of Chandra observations so that a user can investigate which papers have been published from each data set, and conversely, a user can obtain the data sets used in a publication. This functionality supports rich, data-intensive scientific endeavors.

Source: Smithsonian Astrophysical Observatory, 2010

As the project lifecycle ends, the project team often disbands, and details about the data may be lost. Therefore, when data are initially collected, all the metadata should also be captured to include what are now and will be required by primary users at each project lifecycle phase, as well as the anticipated secondary user. These metadata requirements should be planned and implemented by the time data collection begins.

**3.2.5 Preservation (84%) and Transfer of Responsibility (70%)**

Although the project team generally maintains the SDM plan during the project, the full data lifecycle will most likely be longer than the project. The planner must develop a strategy for **preservation** of the data after the project is closed.

Scientific data can become voluminous very quickly. At some point in a project, the data management plan may require migration of copies of some data from a primary working location to another location. This can help the research process by making the current data easier to find, and it may also reduce storage costs. To ensure cost effectiveness, a valuation methodology should be established. Such a

methodology should recognize the uniqueness of the data, as well as the cost of its preservation. The ability to reproduce data and resource availability are other attributes to be evaluated to determine retention time.

The figure below (Figure 3.2.5-1) from the linear lifecycle models (See Figure A2 in Appendix A) illustrates a **transfer of responsibility** for data as part of that lifecycle. Participants discussed that provisions need to be made for this transfer and emphasized the need for institutional repositories supported by appropriate personnel and resources to perform data curation.

In the case where no destination exists for data at project close, planners should develop a plan for maintaining or disposing of the data. For example, the planner may seek a party interested in and capable of managing the data for the benefit of the larger research community. This search may necessitate a long lead-time, requiring the effort to start well before the end of the project, having staff contact discipline-specific professional societies, journals, and other agencies.

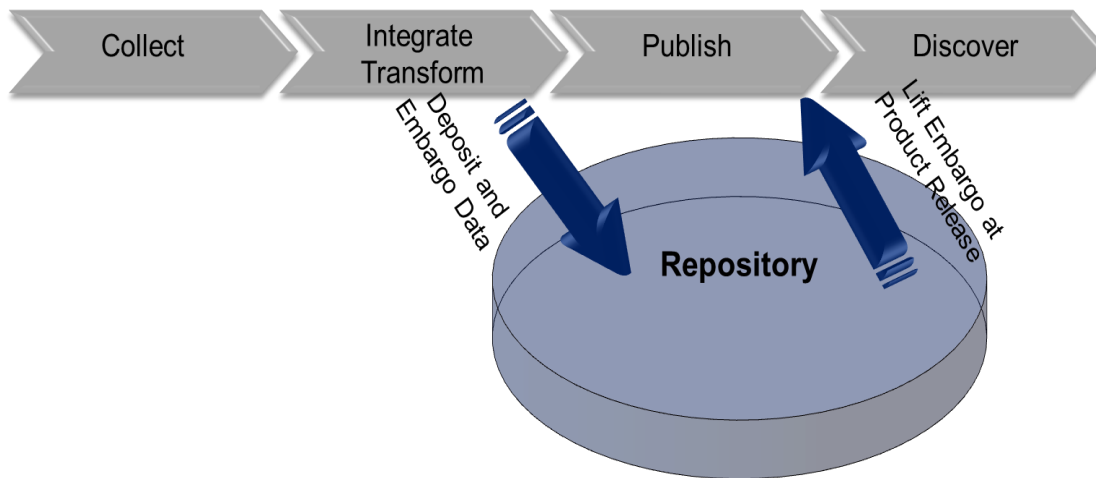


Figure 3.2.5-1. Process to transfer responsibility for data objects.

### 3.2.6 Data Architecture (69%) and Database Operations Management (69%)

A majority of survey participants designated **data architecture** as a required SDM plan element. This view may reflect recognition that federal intramural projects are dependent on the architecture available within agencies. With regard to extramural projects, the survey results may indicate the desire of federal scientists and managers to understand the technical shell that will house data that could be of significant use to them in the future. Regardless of the perspective, there is a need to better integrate SDM with enterprise architecture. The data management aspects of many science projects are handled directly by project personnel or through contract personnel who may not be familiar with agency enterprise architecture strategy and implementation. If data architecture considerations are required elements of the data management plan, project personnel will need to address technology issues during project planning. This process can encourage more communication between the mission and the business operations organizations within agencies.

**Database operations management** received an equal amount of support as data architecture in consideration as a required SDM plan element. This result can be viewed from the perspective of researchers, operational users, and science managers, or from that of data managers. Researchers are interested with varying degrees of urgency in having their data available to them in a technical environment that meets their needs. The data managers, if not the scientists generating the data, are

interested in this element of the data management plan as soon in the project lifecycle as possible so that they understand the operational requirements and are prepared to meet them.

### **3.2.7 Reference and Master Data Management (7%) and Data Warehousing and Business Intelligence (54%)**

The concepts of **reference and master data management** seemed to be largely accepted by survey participants as a required data management plan element. Survey participants showed less enthusiasm for the related technical approach of **data warehousing and business intelligence**. This may reflect the common use of reference data in a science environment, while terminology-like data warehousing and business intelligence are approaches that seem more appropriate for administration and business.

### **3.2.8 Data Quality Management (97%), Provenance (88%), and Usability (61%)**

**Data quality management** and fully understanding the **provenance** of data both received almost universal acceptance from survey participants as required elements of a data management plan. These results suggest a preponderance of interest in secondary users being able to know and evaluate data based on an understanding of the processes applied to the data to ensure their quality. The interest is not necessarily an indicator of quality embedded in the metadata, which would reflect a measure of how well the data meet quality objectives of the project for which they were originally used. Likewise, providing the full provenance of data enables an evaluation of a broader array of factors. In fields such as environmental assessment, where recommendations for policy are often based on data embedded in publications, having access to data provenance through the metadata provides additional defensibility by enhancing potential reproducibility of results.

Survey results indicate that expectations for documentation of extramural research differed from intramural (42%) research. Some of the differences were more stringent data quality, and provenance evaluations for data generated by extramural research.

The concept of addressing **usability** within the data management plan was less well received. This may reflect that addressing usability for the primary project is a trivial exercise, and addressing usability for future projects is difficult to anticipate. This led to the conclusion that documented quality processes and provenance provide the best resource for future data users to use to evaluate data for their intended uses.

### **3.2.9 Value-added Services for the Data (30%) and Workflow Systems (35%)**

Data collected or used by a project are raw material input to the analytical process. In order for analysis to occur, **value-added services** may need to be applied to the data to transform it in a variety of ways. For example, geospatial data may be converted from raster to vector format and integrated with other data. These transformations, integrations, and the analytical process itself produce new data and other forms of digital products. A minority of survey participants indicated that value-added services should be a required element in a data management plan. This may reflect a perceived difficulty in advance planning for transformations and acceptance that, if deemed of future value, the products resulting from these transformations will be documented as digital products in their own right. It may also reflect that documentation of these services should be considered beyond the scope of individual project SDM plans.

Review of products and publications is standard practice in federal science organizations. In the digital environment, these review processes can be greatly expedited by **workflow systems**. One source of recommended SDM plan elements indicates that these systems should be included in an organization's SDM plan. Relatively few survey participants agreed that anticipated use of workflow systems should be included in SDM plans.



## 4.0 RECOMMENDATIONS

As was fully described in Section 1.0, workshop results and recommendations are built on previous work in the scientific data policy and data management planning areas. This includes work performed in an interagency context, as well as the work of many individual agencies and COPs. The workshop added operational viewpoints to the discussion by including input from federal workers that produce, manage, and use scientific data. The recommendations address data management policy and planning, and they suggest ways to ensure that policy and planning are supported by operational realities.

One important outcome of the workshop is the recognition that the value of data as a national asset is not well understood. This argues for a more systematic research and assessment program to better understand how data as an asset is integrated into our economic system and how it impacts both science and society.

### 4.1 Recommendations to Federal Policy

#### ***4.1.1 In Order to Better Manage Scientific Data as an Enterprise Asset, Policy and Research Funding Agencies Should Support Gaining a Better Understanding of the Value Proposition for Effective Scientific Data Management***

When data are discussed as a managed enterprise asset, the concept should not be strictly in the context of an agency balance sheet of assets and liabilities, but rather as something with intrinsic value. However, it is necessary to understand the value proposition and the objective details of the costs and benefits of effective data management. However, value calculations are difficult to define, and the cost-benefit analysis is complex. No systematic method was identified during the workshop as a best practice in assessing either costs or benefits, but rather, case studies and vignettes were cited that demonstrated the benefits of treating data as an enterprise asset. Unfortunately, case studies are often subjective and case-specific. A more objective, quantitative approach is needed for valuation of data to the broader enterprise.

Part of the cost value assessment must address the issue of cost distribution to data producers and curators versus the benefits that accrue to secondary data consumers. Secondary use of data adds to its potential value, but preparation of data to support secondary use is a significant expense. Additional complications occur when data will be maintained beyond the lives of individual projects. Planning approaches are needed to enable complete valuation of data assets, accounting for both primary and potential secondary use of data.

#### ***4.1.2 Agencies Should Consider Portfolio Management as a Model when Determining How to Allocate Resources to Manage and Preserve a Complex Array of Data Generated or Held by an Agency***

The concept of portfolio management is based on balancing multiple factors to create an optimized set of assets to support an enterprise mission. Agency policy should encourage application of portfolio management to science data assets. Management of the data to be produced should be included as a cost factor when projects are submitted for initial investment decisions. As in the management of any asset portfolio, each agency still must develop its goals for return on investment and risk tolerance. All agencies could benefit from some research to determine how to apply concepts of portfolio management to agency data preservation decision making. The capital planning and investment control (CPIC) process is an example of portfolio management as applied to federal information technology (IT) investment. Agency data collections should be reviewed periodically to ensure that resources allocated to preservation are applied for optimum effect. Use of portfolio management techniques to facilitate this process should be explored. Use of such techniques helps evaluators to understanding the value of data sets individually and as part of a managed collection.

### **4.1.3 Agencies Should Stimulate Cultural Change through a System of Incentives to the Stakeholders**

Over \$150B per year is spent on R&D. More than one third of workshop participants indicated that data from that R&D is not managed as an enterprise asset, creating a commitment gap. Resources and culture are the two key impediments that have been identified as barriers to effective data management. This suggests the need for exploration of cultural adjustments to close the gap. Given the economic situation, the cultural route is more likely to produce results. The concept of ownership of scientific data has been changing dramatically as data are recognized as an asset. Policy should help expedite the change in the perception of research teams from being owners of the data they generate in their research to being custodians or stewards of the data to support the agency mission. This situation may engender their perception of losing control of “their” data while being burdened with extra work to serve the needs of others. This represents a high-impact cultural change for agency researchers

It is recommended that the pressure from the top should address issues related to incentives for data sharing, and these considerations should be included in policy and in the support for data management planning.

#### **4.1.3a OPM and Agency Reward Structures Should Reflect Data Policy Objectives**

Policy in and of itself does not create cultural change. It is the complex mix of factors, including implementation of effective rewards structures that will ultimately create the effective policy implementation.

It is important to understand the view of responsibilities through the lens of consumers and producers of data. To effectively treat data as a national asset, data producers need to expend resources to make data broadly sharable. Consumers of data benefit when real effort has been made by producers to document and make data accessible in a manner that provides sufficient context for secondary use. Currently, an incentive structure does not exist to equitably distribute the cost of making data sharable among producers and consumers. The idea of treating data publication in the same manner as publication in scientific literature was discussed as a model to incentivize data stewardship. Providing easy, standard means of citing datasets will ensure credits to individuals and organizations responsible for data and will encourage data sharing.

#### **4.1.3b OPM Should Support Cultural Change through Modifications in Standards for Researcher Promotions and through Establishment of a Series for Data Management and Curation; Training Should be Developed to Leverage Agency Personnel Resources for Effective Data Management**

Governance and stewardship addressed in the data management plan involve allocation of personnel assets to perform data management functions in order to accomplish plan objectives. Approaches discussed in the breakout sessions included training existing science personnel to perform data management functions, training data management personnel in relevant aspects of science (including data management personnel from the outset of the project), and leveraging assets and capabilities in related programs like quality assurance and enterprise architecture.

### **4.1.4 A Solid Interagency Coordination Function Should be Established and Maintained**

Participants recognized that science challenges do not respect agency, domain, or discipline boundaries and that SDM is the critical capability needed to facilitate collaborative effort. Because COPs often cross agency lines, there must be elements to address policies that require interagency consideration. Interagency policy should provide an “umbrella,” supplemented by more detailed policy at the agency,

program, project, and COP level. Many references in these recommendations point to the need for interagency action. This is consistent with the third recommendation from the *Harnessing Report*. While the need for an NSTC subcommittee was not specifically addressed at the workshop, the importance of the IWGDD was discussed, so this is consistent with the intent of transforming the IWGDD into a full subcommittee.

#### ***4.1.4a Federal Coordination Must Also Connect to Other Sectors and International Activities***

A broad array of interest groups is working on different aspects of SDM. Strategies must be developed and implemented to leverage public and private resources in a more coordinated, methodical way to expedite progress in meeting SDM challenges.

#### ***4.1.5 A Scientific Data Research Agenda Should Be Established to Provide an Objective Foundation for Scientific Data Management Decisions.***

There is increased recognition of a growing need to understand the data asset in data-intensive science and in our information economy. This creates a need for a more systematic research agenda, or a supplement to existing research agendas, to provide feedback to data management decision-making and operations communities. Successful implementation of policy initiatives depends upon enhanced understanding that would result from improved, formalized feedback. This challenge should be fully explored.

Valuation of data for unknown but potential secondary use is difficult. Very little helpful guidance exists on balancing cost of preservation and providing access with potential benefit from possible secondary use. Policy should encourage development of a repeatable process for valuation of data.

Guidance on how to evaluate and value science data for long-term preservation should be developed and provided to the federal science and SDM communities. This will require systematic understanding and methods, and it should be part of a scientific data research agenda.

#### ***4.1.6 Open Government Goals Should be Supported by Data Management Policy and Planning***

As the government looks to its plans for Open Government through the development of such tools as Data.gov, it is important to integrate these tools into the overall federal architecture and project lifecycle. Federal objectives of transparency and open access to data can only be met sustainably and economically if they are integrated with the business process of science and supported by an interoperable federal architecture. SDM policies and planning are needed to enable this environment to exist.

## **4.2 Recommendations to Agencies on Data Policy**

### ***4.2.1 Each Agency Should Have a Data Policy that Should be Developed in a Federal Policy Context and Should be Compatible with Programmatic and Community of Practice Policies***

It was clear from the results of the Workshop that having an agency policy was both important and possible. Any agency “umbrella” policy would have to be supplemented in agencies with domain, programmatic and community of practice policies. These would need to be upwardly compatible with the umbrella policy.

#### **4.2.1a Agency Data Policy Should Adopt Guiding Principles that Tie into the Federal Data Policy Context**

The discussions and material from the workshop supported development of a statement of guiding principles for digital scientific data preservation and access. This statement should be included in an agency-level data management policy. The seven guiding principles from the *Harnessing Report* are a good starting point for agency consideration. Additional insight into their implications and impact on operations are provided in the details of the report.

#### **4.2.1b Agency Data Policy Should Ensure that SDM Processes are Integrated with Knowledge Management Initiatives**

Although *knowledge management* is defined differently in different agencies, there is significant value in providing data in context with linkages to other data objects that are related to the data and the project. This includes connecting the SDM environment to administrative environments, including budget and human resources. Use of robust metadata, technical approaches to link related digital objects, and ontologies should be considered as tools.

#### **4.2.1c Agency Policy and Structure of Responsibility Should be Clear Regarding Data Retention**

Not all data need to be kept. Agencies should retain data commensurate with their value, and policy needs to be clear about how this is determined and who would make the decisions about data retention. There is a great variety of perceived decision makers who determine data retention policy. In many cases, these decisions are made by default rather than by systematic process. Participants only had moderate awareness of the NARA role in records retention and whether data were seen as part of the records management process. NARA's role in providing guidance on the valuation of scientific data for retention should be coordinated with the larger research agenda to analyze the costs-benefits of managing and preserving specific scientific data.

#### **4.2.2 Agencies Should Manage Scientific Data for Appropriate Control while Making Appropriate Access More Transparent**

There are problematic barriers to scientific data access and exchange among agencies and with outside COPs. Scientific data access control and data security concerns stem from issues with privacy, confidentiality, intellectual property, and other areas. Researchers and operational users in particular indicated that they had need of other agencies' data and have difficulty gaining access to use it. The landscape of restrictive categories is difficult to navigate, and categories seem to be applied in a non-standardized way across agencies. There is a dearth of long-term repositories available to provide access to science data under a homogeneous classification system. Agencies should consider how they restrict data and should make it easier for other agencies to get to needed data.

Policy should be used to clarify and balance access versus control of data. This closely relates to elements of policy that should catalyze development of incentive structures to reward researchers and agencies for sharing data. Interagency-level work on controlled unclassified information (CUI) should be more widely vetted and understood as it applies to SDM.

#### **4.2.3 Agencies Should Establish the Role of Chief Data Officer and should Clarify Roles and Responsibilities of Agency Personnel**

There is a need for policy to clarify the roles and responsibilities of agency personnel. Implementation of the position of chief data officer (CDO) is a key to deriving full value from science data and data assets. There are concerns related to the real cost to the agency of implementing a CDO position, defining the relationship of the CDO to the CIO, and by extension, the relationship of SDM to enterprise architecture.

An effective policy should address these concerns and should affirm CDO authorities and responsibilities and those of related officials.

Within the organizational structure, there is need for continuously funded data curation, stewardship functions, and staff within agencies to ensure continuity of SDM. Data stewardship should be a responsibility at all levels of an organization. Particular attention should be given to responsibilities for long-term management of data after project close, and when organizational change occurs (such as personnel with key understanding of data moving to other assignments).

Regardless of whether the position of CDO is a new, separate position or whether the role is integrated with the duties of an existing high level official, policy should be implemented to encourage development of the functions of a CDO position and, to the extent possible, the clear and visible assignment of those functions. Policy should address target outcomes for data stewardship that the CDO should achieve.

Although operating structures are very agency dependent, establishing a review committee to help make determinations about data retention and curation should be considered. The issue of stakeholder involvement should also be considered either in policy or in data management planning infrastructure. The interface between data policy and planning when managing organizational structures is critical. Agency policy and planning should be overseen to ensure consistency.

### **4.3 Recommendations to Agencies on Data Management Planning**

#### **4.3.1 *Scientific Data should be Managed According to an SDM Plan that Covers the Full Data Lifecycle and Also Supports the Full Project Lifecycle***

The *Harnessing Report* guiding principle—“Longer preservation, access, and interoperability require management of the full data lifecycle”—was strongly supported by the workshop participants. In addition, policy should require that data management planning be well integrated into project planning. It should begin at project/effort inception and should be an integral part of project planning, budgeting and management. Policy regarding lifecycle planning should also draw from new and emerging technologies in a cost effective way.

Flexibility in policies and data management planning should be balanced with requirements for interoperability, access, and preservation. Policy, procedures and technical solutions can help achieve the balance, but these need to be put into a framework and addressed at a sufficient level of detail.

SDM approaches must be harmonized with changes to approaches in science. Large repositories of data provide opportunities for data mining and “mash-up” analysis that previously did not exist but are data-dependent on appropriate data management approaches.

##### **4.3.1a *Transfer of Responsibility Must Become a Part of the Lifecycle Management Process***

Because the full data lifecycle will most likely be longer than the project, the planner must develop a strategy for preservation of the data after the project is closed. At some point in a project, the data management plan might require migration of some data from a primary working location to another location. This can help the research process by making the current data easier to find, and it also may reduce storage costs. To ensure cost effectiveness, the valuation issue discussed in section 4.1 should be addressed and will need to be supported by the research agenda discussed in section 4.1.5. A valuation methodology should recognize the uniqueness of the data, as well as the cost of its preservation. The ability to reproduce the data and resource availability were other attributes to be evaluated to determine retention time.

The concept of transfer of responsibility has not been generally accepted for data as part of that lifecycle, but it needs to become an integral part of the process.

### **4.3.2 The Data Management Plan Should Be “An Ongoing, Open-ended Living Electronic Record” that Follows the Data through its Lifecycle**

The data management plan should be initiated early in project planning and maintained over the life of the project, including throughout disposition of data at project close-out. Data planning must be incorporated into project planning, and data plans should be linked to other project documentation such as relevant publications and quality assurance plans.

Much of the SDM plan could be drafted by data management personnel and then completed in conjunction with the principal investigator and team. SDM plans need to be adaptable to project changes, and they should be stored in a repository to enable project staff to access them for reference, update, and to make it available for other agencies’ staff members for review. Upon approval, SDM plans should be stored in a location accessible not only to the project members, but also to other project leaders and/or principal investigators looking for examples of SDM techniques.

The data management plan connects and documents the data lifecycle with the project lifecycle. New approaches to data management planning leveraging metadata standards, linking techniques, and repositories for the various types of digital objects created during the project lifecycle should be explored in an effort to improve and expedite data management planning.

## **4.4 Recommendations on COPs**

### **4.4.1 Agencies Should Make Effective Use of COPs**

The *Harnessing Report* principle that, “*Communities of practice (COP) are an essential feature of the digital landscape*”<sup>6</sup> was confirmed during the workshop. Policy should be leveraged to develop and recognize COPs as significant players in effective data management and planning. In addition to enabling sound governance and data stewardship, formalizing COPs is a promising approach to allocating the cost of managing data across a larger group of consumers with interest in the data, potentially relieving individual projects of the full cost of managing their data for secondary as well as primary use.

One COP with a long history in information management that is increasingly involved in management of many types of digital objects is the Federal Library and Information Center Committee (FLICC).<sup>7</sup> Many of the approaches used historically to manage publications are relevant to SDM. There are also data management communities in the context of technical data management and in the CIO context. The SDM COP has different characteristics from all of these, but it can learn and borrow from them. This report relies heavily on input from the CIO community of data management, including the use of the DAMA report.

Agencies should encourage and facilitate planning efforts not only within their organizational structures, but also with COPs within and across federal agencies and private entities.

---

<sup>6</sup> *Harnessing Report*

<sup>7</sup> Two federal groups that share best practices in this community are CENDI (the Federal STI Managers Group) and the Federal Library and Information Center Committee (FLICC).

#### ***4.4.1a An Interagency Coordination Group Should Develop a Conceptual Framework for COPs that Clarifies the Unique Aspects of Scientific Data Compared to Technical and Administrative Data***

There is enormous diversity in the way projects are configured to meet science challenges. The data management approach used to support mapping of the human genome may be quite different than that needed by a principle investigator making seasonal transects of the Chesapeake Bay. These differences dictate that one size of data management plan will not fit all.

A framework is needed for data management planning that agencies can use as a starting point and modify as needed. A fully described set of SDM functions or a model of the overall SDM process could help address the community's uncertainty. For example, there is a relationship between the size, complexity, number of data sets, access requirements, processing requirements, and data preservation needs of projects that drive the scope of the required SDM plan. However, there is currently little guidance available to assist scientists and data managers in determining SDM plan scope. There is a concomitant need for SDM planning training that would be augmented by good examples of SDM plans from particular domains or COPs. Guidance is also needed on how the SDM planning function will interact with other agency-wide programs such as project management, quality assurance, and enterprise architecture. Existing governance processes like those included in the CIO function and enterprise architecture need to be reconciled with SDM governance.

A sample SDM plan table of contents is provided in Appendix D and can be considered as a starting point to be tailored for a diverse set of SDM implementation scenarios.

#### ***4.4.1b Agencies Should Support the Role of COPs through Increased Outreach to Help Primary Data Generators Identify Future Users of Their Data***

The current federal SDM environment is a patchwork of higher level, specific data management policies. There has been some attention paid to data rights in the same scattered fashion. This has contributed to the current inconsistencies and inefficiencies preparing data for users outside the original project team.

### **4.5 Recommendations on Infrastructure**

#### ***4.5.1 Agency Infrastructure Should Support Scientific Data Management***

As agencies increase requirements for data management, it is also important for them to help create or support the infrastructure that will be needed to allow project managers to fulfill the new policy requirements. This will help change the agency culture and practice with regard to SDM. Agencies should evaluate their infrastructures to support data lifecycle management. For example, if an agency provides an institutional repository either directly or through supporting initiatives to its COPs, then multiple projects could incorporate it into their plans for storage and archiving. This aids the project manager in defining this part of the lifecycle, and it helps to gain economies of scale and coordination.

#### ***4.5.2 Policy and Individual Agencies Should Support Persistent Identification across the Government, Including Version Control to Facilitate Data Management and Use***

It is important to creating an environment in which digital objects are linked and version and change control are maintained on datasets and related objects. Operational users in particular need the correct versions of models to their corresponding input parameters and outputs. Where science is conducted in a digital environment, this aspect of data management is critical to legal defensibility of policy decisions as well as scientific reproducibility of results. It also can save time and money to ensure that people are reworking experiments where data has already been updated. Problems exist with accessing the appropriate version of data.

A persistent identifier approach for the federal sector will make linkage of digital objects easier. Although linkage is occurring within some disciplines and COPs, best practices are being adopted slowly. Change control on datasets requires continual effort and may not be feasible once projects end. Federal-wide SDM policy should encourage the use of a uniform approach to persistent digital data set identifiers.

#### ***4.5.3 Data Management Planning In and Across Agencies Should Include a Commitment to Effective Metadata Management***

Metadata requirements needed to satisfy the requirements of varied groups of researchers, operational users, science managers, and policy analysts differ. For example, the rigor of research might require awareness of data coverage, precision, methods, assessment of usability, access restrictions, security required once downloaded, and full provenance of the data, while operational users with thoroughly vetted data in an operational environment may have more interest in those elements that enable rapid discovery and access.

However, metadata has a critical role in enabling discovery, sharing, accessing, understanding, and using science data. The importance of metadata is gaining in appreciation, but it needs to be encouraged further. Linkage of the concepts of preservation and access through the full lifecycle requires increased attention to those elements needed to enable discovery, access, understanding and use. These needs include metadata, tools able to access the metadata and data, and tools used to create, analyze, and model the data. Metadata should be linked directly with data, and technologies should be identified and implemented to enable linking and discovering. Ontologies should be established or identified and implemented to enable and facilitate linking and discovery of data. Metadata should be developed and provided early in the project and data lifecycle. It was recognized that periodic review and appraisal of metadata should occur, and that based on this process metadata could change. It was also recognized that COPs should have the latitude to develop metadata structures with extensions, and that persistent unique identifiers for digital data sets are a critical component of data and metadata.

These are expensive processes. The degree and methods used must be based on the valuation of data. As data policy encourages development of metadata sufficiently robust to support secondary use, the development of appropriate community standards gains import. The role of metadata needs to be better understood and coordinated. Aspects of the metadata challenge have long been the province of archivists and librarians. These resources should be encouraged to work in concert with agencies and COP in order to address this need.



## 5.0 CONCLUSION

Recognition of the importance of our scientific data and the need for attention to its proper management has been growing rapidly. Even since the workshop in June, a number of significant policy actions have attested to this importance. This includes Section 103 of *American Competes Reauthorization Act*, which requires OSTP to establish a working group under the NSTC “with responsibility to coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, including digital data and peer-reviewed scholarly publications.” The President’s Council of Advisors on Science and Technology issued a report in December 2010 entitled “Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology.” This report points to the cross-cutting theme of the growth of data volumes. The report recommends observes and recommends the following: “The collection, management, and analysis of data is a fast-growing concern to NIT research. Automated analysis techniques such as data mining and machine learning facilitate the transformation of data into knowledge, and of knowledge into action. Federal agency needs to have a ‘big data’ strategy.” Finally, the President’s Science Advisor’s memo: “Scientific Integrity Memo,” issued December 17, 2010, assigns a significant role to dissemination of scientific information. It states that agencies should develop policies that

*Facilitate the free flow of scientific and technological information, consistent with privacy and classification standards. Open communication among scientists and engineers and between these experts and the public, accelerates scientific and technological advancement, strengthens the economy, educates the Nation, and enhances democracy. Consistent with the Administration’s Open Government Initiative, agencies should expand and promote access to scientific and technological information by making it available online in open formats. Where appropriate, this should include data and models underlying regulatory proposals and policy decisions.*

Activities are also advancing abroad and in other sectors of U.S. science. It is of great importance that the federal dialog be continued and that it also expands to embrace the best ideas and practices from the rest of the science policy communities. The recommendation of the *Harnessing Report*, which was further endorsed by this workshop, combined with the requirements of *America Competes Reauthorization Act*, provides a continuing focal point for management and preservation of U.S scientific assets. The efforts of the SDM workshop participants on work products, findings, and recommendations can form part of the body of resource knowledge for future developments.

To the IWGDD sponsor, the information presented can provide input for a continued federal agenda. For our CENDI sponsor, the information from the workshop can be helpful as each agency continues to development its data policies and plans. For our EPA sponsor, we offer the findings and recommendations of this workshop as input as EPA takes a leadership role as an agency committed to developing and managing its data resources to support both science and regulation.

## APPENDIX A

## APPENDIX A: Community of Practice (COP) Lifecycle Models

During the workshop, it was noted that many communities of practice (COPs) have developed their own lifecycle models that take into account special needs or interests of their community.

### Federal Geographic Data Committee (FGDC) Lifecycle

Figure A1 below shows the FGDC lifecycle model, which advocates compliance of Office of Management and Budget (OMB) Circular A-16, "Coordination of Geographic Information and Related Spatial Data Activities." This framework encourages "timely and high-quality geospatial data to support business processes and operations; stronger partnerships across all levels of government and, when appropriate, the private sector, to increase cost efficiency and return on investment; and improve strategies for completing and maintaining nationally significant themes and datasets associated with OMB Circular A-16 to enhance service to citizens" (FGDC, 2010).

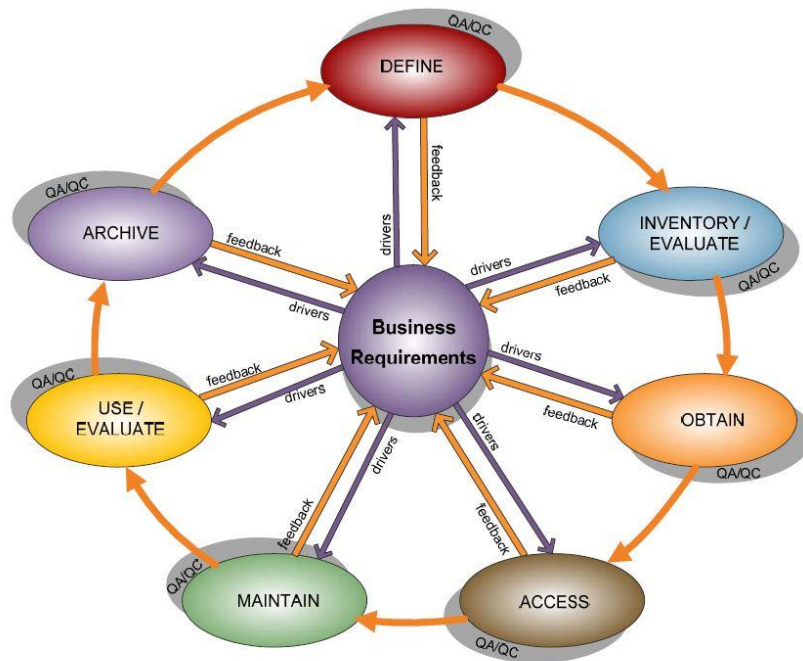


Figure A1. The FGDC data lifecycle (FGDC, 2010).

## Linear Lifecycle

A linear lifecycle model was suggested as being easier to work with in an operational environment. Figure A2 shows a linear lifecycle adapted for use. This model was developed by the Environmental Protection Agency (EPA) Office of Research and Development (ORD). The figure highlights the importance of governance and communications as key aspects of implementing a lifecycle approach. Element processes are defined below.

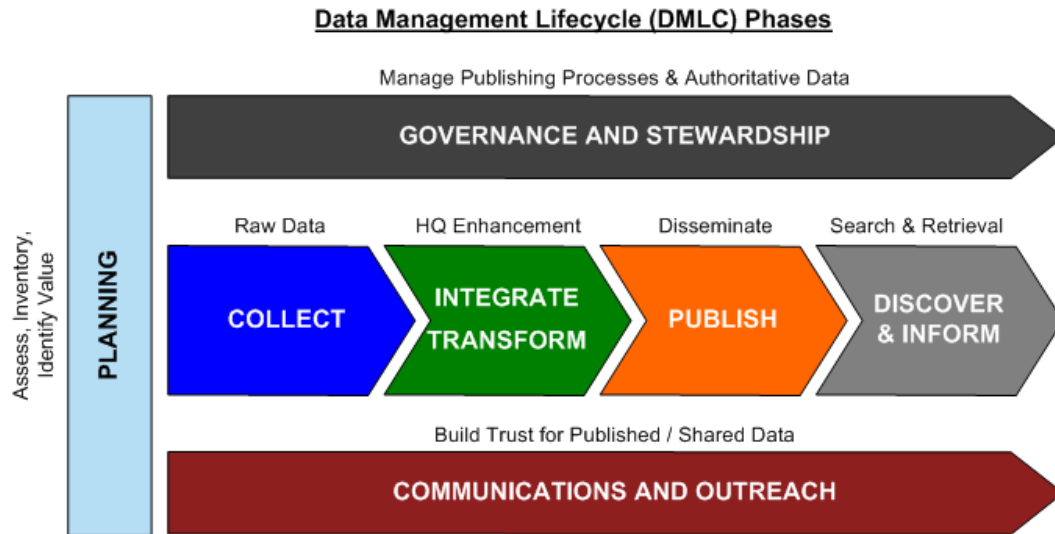


Figure A2. A linear data lifecycle for use in an operational environment (EPA, 2010b).

1. **Plan:** Data are assessed and inventoried by open government governance bodies and segment architects, and high-value sets are identified for sharing with the public through open government initiatives such as Data.gov.
2. **Collect:** Data are collected by the source entity, source providers push data to federal source systems, which provide data in a specified file format, and then deltas are managed by source record tags in the specified file format.
3. **Integrate and Transform:** The agencies integrate the raw data that were collected and add value through various means, including input from individual program offices and scientific research projects. The data are transformed from their initial state and stored in a value-added state, such as through web services.
4. **Publish:** Information resources are prepared for publishing to one or more of the many audiences, including congress, the public, tribal governments, academia, research and scientific partners in non-governmental organizations, other federal agencies, and other stakeholders (e.g., industry, communities, researchers, the media, and Data.gov audiences).
5. **Discovery:** Agencies manage search and retrieval for various internal and external audiences. Discovery will become more complex as secondary audiences are supported through open government initiatives. Secondary audiences need to be informed of the meaning of data as understood by primary audiences, who are more familiar with the environmental legal landscape.
6. **Governance and Stewardship:** This element of the process defines governance bodies and agendas, and it gains acceptance of data steward roles. Governance and stewardship manage the publishing process for ongoing change control, and they maintain versions of the truth across common data.
7. **Communications and Outreach:** This aspect provides for inventory of high value data sets, enables technologies and controlled vocabularies for re-use, allows for management of information exchange agreements, and encourage re-use through communications.

## APPENDIX B

## APPENDIX B: Policy Elements and Principles

Figure B1. Policy elements and principles.	Goals & Principles	Roles & Responsibilities	Best Practices	Issues
<p><b>Manage Scientific Data as Enterprise Assets or Liabilities (EPA)</b>            Scientific data developed with ORD resources (such as funding, staff, computers and other equipment) belong to the taxpayer and are governed by EPA. These data have value, which may be positive (i.e., assets) or negative (i.e., liabilities). An example of an asset is a data set that will be reused for trend analysis; an example of a liability is a data set that will never be reused but ORD continues to incur costs for its maintenance.</p>				
<p><b>Develop a Scientific Data Management Plan that Covers the Full Data Lifecycle (EPA)</b>            Developing a SDM plan provides the opportunity to focus on scientific data, including how the data will be gathered, processed, and analyzed. This plan can involve all stakeholders (e.g., users and potential users of the data) who can assess the value of the data for both current projects and potential future uses and reuses, even beyond the life of the projects. For ORD projects, a data management plan is sometimes developed as part of the quality assurance (QA) plan.</p>				
<p><b>Identify Scientific Data with Metadata to Enable Needed Business Operations (EPA)</b>            To gain maximum value from scientific data, it must be easily accessed and understood by those who use it. The information that provides this understanding is “data about scientific data” – metadata. Metadata can provide “provenance” or data lineage (e.g., by linking to information products such as the data management plan or final report) and can also enable data discovery, retrieval, and appropriate reuse. Metadata is essential for identifying, searching for, locating, storing and retrieving scientific data. The consistent development and use of metadata enables communication between cooperating agencies and the public users of data, and can be used to identify appropriate data users and help provide access control.</p>				
<p><b>Manage Scientific Data for Appropriate Control (EPA)</b>            Development of this policy area will include, among other things, guidance on understanding data rights and circumstances (e.g., proprietary data) that create different types of data rights, policies to establish and maintain an identification process for IP, and guidance on establishing levels of control and how to select the appropriate level of control for a data set given specific data rights.</p>				
<p><b>Maintain Version and Change Control on Data Sets (EPA)</b>            Control of scientific data is needed to ensure the integrity of the data and the final product. Data within a project undergoes a continued development phase, from working data to mature, released, submitted, and archived data. This includes, for example, developing naming conventions and other approaches to maintain version and change control. Not all data require the same level of control, depending on customer-imposed requirements, and agency requirements. One factor to be considered is the maturity of the data set. For example, putting the data under control too early in its lifecycle becomes burdensome and yields little business value. Control of data within a project might be considered as important as the control of the final product.</p>				
<p><b>Retain Data Commensurate with Its Value (EPA)</b>            Data should only be retained as long as it has value to current or future users. There must be a method in place to ensure adequate retention and preservation of data that have value to the agency and how to dispose of data that no longer have value. Data can be retained in many ways, at differing costs (e.g., on-line, near-on-line, archives). Determining the probability and value of future use and the appropriate retention mechanism and timing requires cost-effectiveness assessment and the participation of all stakeholders, including those who represent potential future users of the data (e.g., librarians).</p>				

Figure B1. Policy elements and principles.	Goals & Principles	Roles & Responsibilities	Best Practices	Issues
<p><b>Ensure that scientific data management (SDM) processes are integrated with knowledge management initiative (EPA).</b> Data management and KM are interdependent. The foundation provided by SDM can enable both knowledge sharing (through discovery and retrieval of scientific data, for example) and knowledge retention by supporting knowledge harvesting when a principal investigator retires or leaves EPA. In all SDM activities, one must remain knowledgeable about KM initiatives.</p>				
<p><b>Statement of guiding principles for digital scientific data preservation and access (IWGDD).</b> The principles should provide clear guidelines for those conducting the data planning and implementation activities of the agency and for those seeking to partner with the agency in pursuing shared data goals. This includes criteria for determining whether data are appropriate for preservation and access. Further, the principles must be in accordance with the provisions of the <i>Paperwork Reduction Act</i> (44 U.S.C. 3501 et seq.), OMB Circular A-130, the <i>America COMPETES Act</i>, the <i>Data Quality Act</i>, the <i>Federal Funding Accountability and Transparency Act</i> (FFATA), and other applicable policy, regulatory, and statutory requirements. The agency digital data policy should cite the relevant governing documents wherever appropriate.</p>				
<p><b>Assignment of responsibilities (IWGDD).</b> The roles of agency offices and officials in implementing the agency digital data policy should be described to ensure clear lines of authority and accountability and to provide transparency for those working within and outside the agency on digital data matters. This should include provisions for a designated, cognizant senior science official serving as science data officer to coordinate the digital data activities of the agency and to serve as representative to the Subcommittee on Digital Scientific Data.</p>				
<p><b>Description of mechanisms for access to specialized data policies (IWGDD).</b> Agencies may support various communities of practice (COPs) and distinct data types, formats, and contexts, and they may have differing programmatic goals, needs, and resources. Such agencies should have a harmonized suite of corresponding, specialized data policies. The comprehensive agency digital data policy should describe mechanisms to provide easy and transparent access to the agency's full portfolio of specialized data policies.</p>				
<p><b>Statement of intentions and mechanisms for cooperation, coordination, and partnerships (IWGDD).</b> The agency digital data policy should describe the agency's intentions and mechanisms for cooperation, coordination, and partnerships across sectors. Such sectors can include government at the national, state, or local levels, as well as industry, academia, education, non-profits, and international entities.</p>				
<p><b>Provisions for updating and revisions (IWGDD).</b> The agency digital data policy must be a living document if it is to remain relevant and effective in a dynamic landscape. The policy should describe the mechanisms to be used for updating and revising the document to ensure it is responsive to change and opportunity.</p>				

**Figure B1. Elements of policy and definitions collated from *Harnessing Report* and the *EPA Survey*.**

## APPENDIX C



# APPENDIX C: SAMPLE Data Policy for Digital Science Data Based on Federal and Industry Best Practices

## 1.0 PURPOSE AND INTENT

Digital technologies are reshaping the practice of science. Digital imaging, sensors, analytical instrumentation and other technologies are becoming increasingly central to experimental and observational research in all areas of science. It is now possible for scientists to share not only their research results, but also the data which the results are based.

The purpose of this document is to establish a governing policy for the management of AGENCY scientific digital data as an agency asset. It is AGENCY's intent, that all research data supported with public funds are made available to the public to the extent possible. This policy will promote the following objectives:

- Ensure reliable preservation and effective access to AGENCY digital data for research, development, and education in science, technology, and engineering.
- Implement AGENCY's commitment to data sharing, promoting secondary data use, and supporting transparency, openness, and collaboration in government with regard to environmental information.
- Improve the quality, availability, currency, and accessibility of AGENCY data holdings, especially those that are shared within AGENCY, across government, and with the public.
- Support governance of, and best practices for, data management across the AGENCY.
- Increase productivity in AGENCY information collection and processing activities as the understanding and use of available data increases.

## 2.0 SCOPE AND APPLICABILITY

Digital science data include any information that can be stored digitally and accessed electronically, with a focus specifically on scientific information used by the federal government to address national needs or derived from research and development funded by the federal government.

The scope of the proposed AGENCY policy addresses all scientific data collected and/or generated by AGENCY or by AGENCY's contractors, grantees, and other partners. AGENCY "Scientific data" is defined as:

Data that AGENCY scientists generate, including:

- Field data
- Lab data
- Models
- Model parameter sets
- Model outputs
- Other data (e.g., quality control samples, sample ID data, and instrument calibration data)
- Data that AGENCY scientists collect from secondary sources (typically referred to as secondary or outside data).
- Data that non- AGENCY personnel generate for AGENCY (i.e., extramural data) through:
  - Contracts
  - Grants
  - Cooperative Research and Development Agreements (CRADAs)

### 3.0 AUDIENCE

This AGENCY policy applies to all employees who are involved in scientific data management and use activities, including headquarters, labs, centers, offices, divisions, and branches. These activities include all aspects of the planning, creation, use, communication, retention and assessment of scientific data. It also applies to non-AGENCY organizations (e.g., states, tribes, localities, regulated parties, volunteer organizations, contractors, cooperative agreement holders, grantees, other federal government agencies, intergovernmental organizations, educational institutions) that design, develop directly or indirectly, compile, operate, or maintain AGENCY scientific data.

### 4.0 BACKGROUND

Increases in computational capacity and capability drive more powerful modeling, simulation, and analysis to link theory and experimentation and extend the reach of science. Improvements in network capacity and capability continually increase access to information, instrumentation, and colleagues around the globe. Digital data are the common thread linking these powerful trends in science.

In December 2006, the National Science and Technology Council (NSTC) of the Committee on Science established the Interagency Working Group on Digital Data (IWGDD). AGENCY is a member of NSTC Working Group. The charge to the working group was to develop and promote the implementation of a strategic plan for the federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering. The group produced the report entitled “Harnessing the Power of Digital Data for Science and Society,” January 2009, which provides the philosophical guidance and guiding principles for the development of AGENCY scientific digital data policy enumerated in Section 5.

AGENCY is committed to establishing a scientific digital data policy meeting the charge of the IWGDD. This policy establishes a consistent organizational approach for managing scientific data. The policy ensures that scientific data collected and/or generated by employees, contractors, grantees, cooperators, and other partners, meet the following criteria:

- Data are protected and preserved over the life of the data and the life of the project(s) and program(s) that use the data.
- The data provide maximum value to AGENCY and other users of the data.
- The data are available for effective, timely access, use, and reuse by AGENCY and non-AGENCY scientists (as appropriate) for research and development, decision-making, and other uses.

Related benefits of a coordinated scientific data management policy include the following:

- Improved data quality
- Improved efficiency and effectiveness of data storage, maintenance, retention, and disposal.
- Increased productivity in AGENCY information collection and processing activities as the understanding and use of available data increases
- Other considerations driving the AGENCY digital data policy include (Reference any AGENCY-specific data and/or enterprise architecture requirements pertaining to the collection, acquisition, processing, documentation, storage, access, maintenance, and retirement of data)

### 5.0 AUTHORITY

There are many statutes and executive orders that impact the management of scientific data. Following are key guidance documents: [AGENCY should add or customize the list.]

The *Paperwork Reduction Act* (44 USC 35) has as one of its key purposes to “ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the federal government.”

The Office of Management and Budget (OMB) Circular A-130, “Management of Federal Information Resources,” specifies that “the open and efficient exchange of scientific and technical government information... fosters excellence in scientific research and effective use of federal research and development funds.”

Section 515 of Public Law 106-554, known as the *Information Quality Act*, required OMB to promulgate guidance to agencies ensuring the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies. OMB's government-wide guidelines, published as interim final on September 28, 2001 (66 F.R. 49718) and finalized on February 22, 2002 (67 F.R. 8452), can be found on OIRA's website. Federal agencies were also required by Section 515 to publish their own agency specific guidelines.

The 1991 Supreme Court ruling in *Feist Publications, Inc. v. Rural Telephone Service Co.* (499 U.S. 340) establishes that “facts do not owe their origin to an act of authorship, they are not original, and thus are not copyrightable.”

Copyright law (17 USC 105) provides that “Copyright protection under this title is not available for any work of the United States Government.”

The *Freedom of Information Act* (FOIA; 5 USC 552) provides for public access to the records of the federal government.

This scientific digital data policy does not rescind any other AGENCY data policy or guidance. Digital data policy is a subset of AGENCY data policy and conforms to existing data policy supported by the existing AGENCY authorities.

[Authority should be the context in which the agency creates data as well as federal laws and regulations that support the policy document.]

Link to NSTC Committee on Science Charter:

<http://www.whitehouse.gov/sites/default/files/microsites/ostp/committee-on-science-charter.pdf>

## 6.0 GUIDING PRINCIPLES

The following guiding principles were articulated by the IWGDD and are accepted by AGENCY as the basis of the AGENCY science data management policy:

- Science is global and thrives in the digital dimension (“The Harnessing Report”).
- Data are national and global assets (“The Harnessing Report”).
- Preservation is a government and private sector responsibility and benefits society as a whole (“The Harnessing Report”).
- Communities of practice are an essential feature of the digital landscape (“The Harnessing Report”).
- Long-term preservation, access, and interoperability require full lifecycle management (“The Harnessing Report”).
- Not all digital data need to be preserved, and not all preserved data need to be kept indefinitely (“The Harnessing Report”).
- Dynamic strategies are required (“The Harnessing Report”).
- Managing AGENCY digital data is the responsibility of AGENCY and is not owned by any individual or business unit.

- Digital data (both structured and unstructured) and the metadata about that data are business and technical resources owned by the public and managed by the AGENCY.
- AGENCY will plan and follow data acquisition policies that ensure the collection of long-term data sets needed to satisfy mission requirements (Adapted from NASA)
- All public-facing datasets accessed through AGENCY are confined to public information and must not contain National Security information (Adapted from Data.gov)
- All information accessed through AGENCY is in compliance with the required confidentiality, integrity, and availability controls mandated by Federal Information Processing Standard (FIPS) 199 (Adapted from Data.gov).
- All information accessed through Data.gov is subject to the Information Quality Act (P.L. 106-554) (Adapted from Data.gov).

## 7.0 POLICY

## 8.0 SCIENTIFIC DIGITAL DATA WILL BE MANAGED AS AN AGENCY ASSET

AGENCY commits to managing scientific digital data developed with AGENCY funds, developed on behalf of AGENCY, or assigned to the AGENCY, as an asset for which it has responsibility as a public trust. It is AGENCY’s responsibility to maintain, preserve, and provide access to this data for as long as the agency determines this to be in the public interest.

AGENCY recognizes that its science data management procedures, standards, and guidelines must be based on a unifying approach – the interdependence of a science project lifecycle, complete with its critical Science data management plan, its data quality plan and descriptive products within a broader context of and Science Data Lifecycle, which are all separate but interdependent. The Science Project Lifecycle of AGENCY is illustrated in Figure C1.

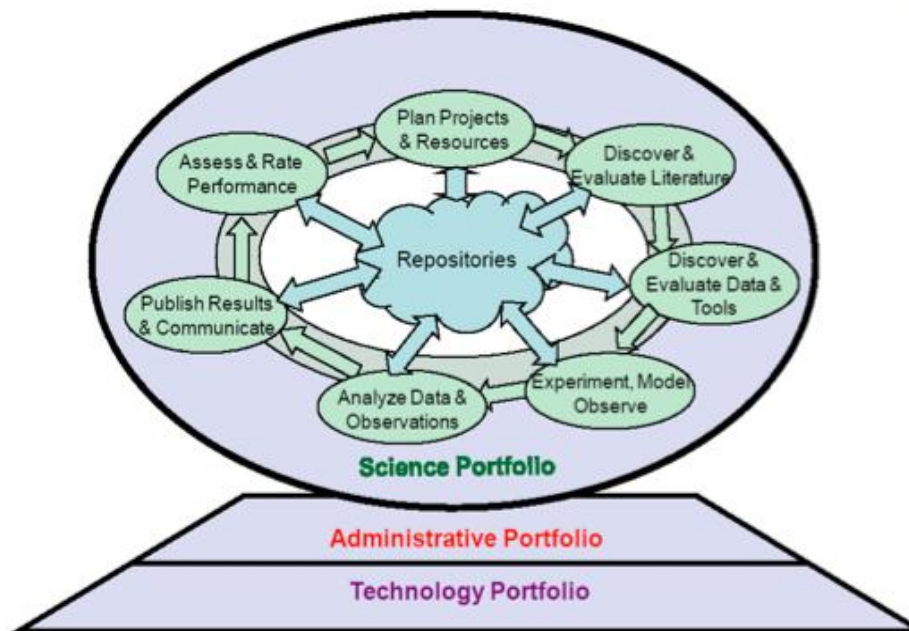


Figure C1. The science project lifecycle.

The science data lifecycle is given in Figure C2. The data outputs of a science project are freestanding artifacts that must be maintained intact, secure, and accessible for future uses, foreseen and unforeseen, but archived or disposed of when no longer useful. Within this linear construct, many data maintenance feedback loops may be interpreted through this simplified model. Programs will plan next set of data acquisitions based on discoveries from the current one and would use lessons learned from data management to plan the evolution of data system for future datasets.



**Figure C2. The generic science data lifecycle.**

AGENCY’s science data management strategy is therefore based on project-specific science data management plans, which are “living documents,” written and updated continuously, that define the management requirements at each stage of the data lifecycle.

To manage and link these two lifecycles, AGENCY recognizes the need to maintain appropriate science data registries to provide access to science data management plans, descriptive information on science data sets, and data sets themselves.

**Example Key Performance Indicators**

- What is the number of identified digital data assets of the AGENCY?
- Is there an AGENCY enterprise data management framework established for the AGENCY?
- Does the AGENCY have an implemented data policy for scientific data?

## 8.1 Establishment of an Oversight Function to Manage Scientific Digital Data Produced with Agency Resources

The purpose of the AGENCY governance is to ensure that AGENCY science data management supports agency mission and the needs of the science community.

AGENCY will ensure that there is governance to oversee scientific data management policy enforcement.

AGENCY science data governance shall be based on an interdisciplinary membership, including representatives from a wide array of AGENCY data stakeholders.

**Examples of Key Performance Indicators**

- Is there an Agency-wide oversight organization focused of the data assets of the AGENCY?
- Is the data oversight organization including representation from all data stakeholders in the AGENCY?
- Is there a Charter established and in use by the AGENCY?

## 8.2 Development of Science Data Management Plans

AGENCY hereby adopts the core findings for data management plans from the Workshop to Improve SDM and will adapt this template to a comprehensive AGENCY version.

Scientific data management plans will be created at the outset of every AGENCY-funded project to prior to initial funding. Scientific data management plans will be registered with the appropriate science data registry under procedures developed by the SDMC. The SDMC will describe appropriate mechanisms to update scientific data management plans to ensure that they evolve and are updated throughout the project lifecycle to reflect project needs and progress.

### 8.3 Agency Science Data Registries

The AGENCY will establish one or more science data registries to provide authoritative information on all science data sets under agency management and related AGENCY science projects. These registries will include links to relevant financial and performance management systems, such as grants management and contract management databases.

Science data registries will support the following functions:

- Preserve and maintain metadata for all AGENCY science data sets, ensuring that metadata is captured throughout the data lifecycle, and made available to AGENCY employees, science researchers inside and outside AGENCY, and the general public.
- Provide links and references to data set locations, so that AGENCY data sets can be discovered, viewed, and retrieved by qualified researchers, other agencies, and the general public.
- Provide science data taxonomies for cataloging and retrieval of science data sets as jointly developed by AGENCY and other research enterprises, including federal, academic, and private groups
- Provide links to Data Quality Plans associated with data assets
- Provides links to grants and funding related to support AGENCY’s science data strategic goals
- Provide repositories, where necessary, for the storage and maintenance of AGENCY data sets that would not otherwise be conveniently accessible to qualified researchers, other agencies, and the general public
- Link AGENCY data sets to the science research projects, external or AGENCY-funded, that developed them

**Examples of Key Performance Indicators**

- Does each data asset have a corresponding data management plan?
- Does each data asset have a corresponding data project plan?
- Are data quality objectives readily available for each data asset?

The AGENCY Science Data Management Council is responsible for the identification and design of all SDRs. The SDMC will also create appropriate and necessary standard operating procedures and compliance requirements for the use of SDRs by AGENCY-funded science projects. The SDMC will appoint managers and stewards of SDRs to ensure uninterrupted oversight, control, and access by authorized parties.

**Examples of Key Performance Indicators**

- To what extent are metadata available for AGENCY data assets?
- Does the AGENCY maintain a central or federated metadata repository?
- To what extent are data assets available as linked open data?

### 9.0 RELATED DOCUMENTS

Related documents include AGENCY data management procedures, adopted data standards, and relevant federal executive orders and directives for cross-agency data management functions.

### 10.0 ROLES AND RESPONSIBILITIES

The AGENCY assigns overall executive responsibility for scientific data policy and data management planning to a chief scientific data officer (CDO). The CDO reports to [AGENCY to complete.] The CDO responsibilities include the following:

- Coordinates execution of data management processes across the lifecycle of AGENCY project.
- Chairs the Science Data Management Council [if AGENCY forms one].
- Oversees the development, access to, and preservation of science data management plans.
- Ensures that science data sets are maintained in accordance with AGENCY data standards.
- Ensures effective operation and maintenance of the science data registries.
- Responsible for ensuring those granting and contracting organizations have controls so that principal investigators comply with science data management plans as a condition of project funding.
- Supports principal investigators in complying with the provisions of this policy.

## 11.0 DEFINITIONS

[Provide definitions of all key terms, acronyms, and organizations referenced in policy.]

## 12.0 RELATED PROCEDURES, STANDARDS AND GUIDANCE

*Information Quality Act*, Section 515 of the *Treasury and Government Appropriations Act of 2001* (PL106-554, 31 USC 3516) (<http://www.fws.gov/informationquality/section515.html>)

M-10-06, *Memorandum for the Heads of Executive Departments And Agencies – Open Government Directive*, Executive Office of the President, Office of Management and Budget, December 8, 2009 ([http://www.whitehouse.gov/omb/assets/memoranda\\_2010/m10-06.pdf](http://www.whitehouse.gov/omb/assets/memoranda_2010/m10-06.pdf))

OMB Circular No. A-130 – “Management of Federal Information Resources,” Appendix III to OMB Circular No. A-130: *Security of Federal Automated Information Resources* ([http://www.whitehouse.gov/omb/circulars\\_a130\\_a130trans4/](http://www.whitehouse.gov/omb/circulars_a130_a130trans4/))

*Clinger-Cohen Act of 1996* (PL 104-106)

*The Paperwork Reduction Act* (44 USC 35)

*The Office of Management and Budget (OMB) Circular A-130*

*The Freedom of Information Act (FOIA)*; 5 USC 552)

NIST Special Publication 800-53: *Recommended Security Controls for Federal Information Systems*, U.S. Dept. of Commerce, December 2007

## 13.0 MATERIAL SUPERSEDED

Initial Release

## 14.0 ADDITIONAL INFORMATION

<http://www.digitalpreservation.gov/>

<http://www.planets-project.eu/>

## APPENDIX D



## APPENDIX D: Data Management Plan Template and Table of Contents

PROJECT XXX Data Management Plan  
Version 2.3

Latest version available at <http://agency.gov/program/projectxxx/DMplan.pdf>  
Revision Date: August 16, 2009

Prepared for the  
ABC Sciences Division  
Office of Biological and Environmental Research  
U.S. Department of FGH  
Budget Activity Number DRD 04 15 52 0

Prepared by the  
Organization Name  
AAA BBB Laboratory  
AnyTown, AnyState, 99999

Author 1, author 2, author 3, etc.

Unique Publication Number yyy-zzz

## TABLE OF CONTENTS

SUMMARY OF CHANGES BETWEEN VERSIONS OF THIS PLAN .....	D-4
ABBREVIATIONS .....	D-5
1.0 INTRODUCTION .....	D-6
1.1 Background .....	D-6
1.2 Project Description.....	D-6
1.3 Data Management Overview.....	D-6
1.4 Data Governance.....	D-6
1.4.1 Data Management Organization, Roles and Responsibilities .....	D-6
1.4.2 Data Architecture Management.....	D-6
1.5 Organization of this Data Management Plan.....	D-6
2.0 DATA ACQUISITION .....	D-6
2.1 Data Quality Objectives .....	D-6
2.2 Data Acquisition.....	D-7
2.2.1 Identification of Data .....	D-7
2.2.2 Applicable data rights and/or access controls.....	D-7
2.2.3 Control of Erroneous Data.....	D-7
2.2.4 Changes to Data (Due to Processing or Other Reasons) and Versioning.....	D-7
2.2.5 Change Control of the Data.....	D-7
2.2.6 Data Validation.....	D-7
2.2.7 Expressing the Amount of a Substance.....	D-7
2.2.8 Time Averaging .....	D-7
2.2.9 Qualifying Data .....	D-7
3.0 DATA MANAGEMENT.....	D-7
3.1 Types of Data to be Collected, Processed, and Utilized.....	D-8
3.2 Data Sources.....	D-8
3.3 Data Management Resources Needed .....	D-8
3.4 Data Acquisition Activities .....	D-8
3.5 Data Storage Activities.....	D-9
3.6 Data Evaluation and Processing Activities .....	D-9
3.7 Time Integrated Data .....	D-9
3.8 Date and Time Formats .....	D-9
3.9 Reporting Missing Data.....	D-9
3.10 Reporting Calibration Values and Uncertainty Estimates .....	D-9
3.11 Data Flags .....	D-9
3.12 Data Access Provisions .....	D-10
4.0 DOCUMENTING DATA .....	D-10
4.1 PROJECT XXX Data and Information Categories.....	D-10
4.2 PROJECT XXX Data Format and Codes .....	D-10
4.3 Metadata Tables.....	D-10

4.4 Measurement Data Table ..... D-10

4.5 Data Formatting and Validation Activities ..... D-11

4.6 Data Formatting Activities ..... D-11

4.7 Data Validation Activities ..... D-11

4.8 Reference and Master Data Management ..... D-11

5.0 ARCHIVAL AND DISSEMINATION OF PROJECT XXX DATA AND INFORMATION .... D-11

5.1 Data Archiving Activities ..... D-11

    5.1.1 Transfer of Responsibility ..... D-11

5.2 PROJECT XXX Permanent Database Archive ..... D-11

5.3 Data and Research Product Access and Distribution ..... D-11

6.0 RECORDS MANAGEMENT ..... D-12

6.1 Records Management System ..... D-12

6.2 Records Identification, Authentication and Indexing ..... D-12

6.3 Records Distribution and Storage ..... D-12

6.4 Records Retrieval ..... D-12

6.5 Records Retention Reviews ..... D-12

    6.5.1 Assessing continuing impact or value of the data ..... D-12

    6.5.2 Determining “Uniqueness” of the Data Collection ..... D-12

    6.5.3 Risks Associated with Retention ..... D-12

7.0 SOFTWARE CONFIGURATION CONTROL AND DATA MANAGEMENT

SYSTEM ADMINISTRATION GUIDELINES ..... D-12

7.1 Project-specific Database and Software Requirements ..... D-12

7.2 Day-to-day Operation of Data Management Systems ..... D-12

## SUMMARY OF CHANGES BETWEEN VERSIONS OF THIS PLAN

Status (Baseline/ Revision/ Canceled)	Document Revision	Effective Date	Description
Baseline	1.0	1/10/2007	
Revision	2.0	4/14/2008	Removed references and rearranged Section 7 to match order of quality records.
Revision	2.2	5/15/2009	Removed section 6, merged content into other sections, and renumbered.
Revision	2.3	6/20/2010	Revised procedures to reflect streamlined process.

## ABBREVIATIONS

AAQS Standard	Ambient Air Quality	FWDR	Far Western Development Region
ADB	Asian Development Bank	GDP	Gross Domestic Product
AQM	Air Quality Management	GJ	Gig Joule
ARI Infection	Acute Respiratory	GR	Graveled Road
BT	Black Topped	GRI Institute	Global Resources
CAI-Asia	Clean Air Initiative – Asia	HC	Hydrocarbon
CAM	Metropolitan Environmental Commission	HCHO	Formaldehyde
CARB Board	California Air Resources	IEE	Initial Environmental Examination
CBS Statistics	Central Bureau of	kPa	Kilo Pascal
CDR Region	Central Development	LPG	Liquefied Petroleum Gas
CNG	Compressed Natural Gas	µg/m <sup>3</sup> Meter	Microgram Per Cubic
CO	Carbon Monoxide	Mn	Manganese
COPD	Chronic Obstructive Pulmonary Disease	MWDR	Mid-Western Development Region
DAD	Department of Agriculture Development	NAAQS	National Ambient Air Quality Standards
DALY Years	Disease Adjusted Life	NGO Organization	Nongovernment
DOH	Department of Health	NO <sub>x</sub>	Oxides Of Nitrogen
EDR Region	Eastern Development	NPC Commission	National Planning
EIA	Environmental Impact Assessment	O <sub>3</sub>	Ozone
EMPC	Environmental Management and Promotion Center	P	Phosphorous
EPC Council	Environment Protection	PAH	Polycyclic Aromatic Hydrocarbons
ER	Earthen Road	Pb	Lead
ESPS	Environmental Sector Program Support	PM <sub>10</sub>	Particular Matter Smaller Than 10-Micrometer Diameter
EV	Electrical Vehicle	PM <sub>2.5</sub>	Particulate Matter Smaller Than 2.5 Micrometer Diameter
		PPM p	Parts Per Million
		RSPM Matter	Respirable Particulate

RVP	Reid Vapor Pressure	TSP	Total Suspended
S	Sulfur	Particulate	
SKO	Kerosene	VOC	Volatile Organic
SO2	Sulfur Dioxide	Compound	
km <sup>2</sup>	Square Kilometer	WHO	World Health
THC	Total Hydrocarbon	Organization	

## 1.0 INTRODUCTION

### 1.1 Background

[This section provides a brief, high-level description of the project, the digital scientific data to be produced, its uses, and the data management plan.]

### 1.2 Project Description

[This section includes a brief description of the research project, its scope and objectives, to provide context for the scientific data to be managed. The project description should include a link or definitive reference to the project plan or other document that describes the project.]

### 1.3 Data Management Overview

[This section provides a brief, high-level description of the digital scientific data to be produced, its uses, and how it will be managed.]

### 1.4 Data Governance

[This section provides high-level planning and control over data management. This section also describes how authority, control and shared decision-making—including planning, monitoring and enforcement—are exercised for data assets.]

#### 1.4.1 Data Management Organization, Roles and Responsibilities

[This section describes the specific implementation of data management for this project]

#### 1.4.2 Data Architecture Management

[This section describes how the project data integrates with and leverages the enterprise data architecture, and how the project’s data architecture connects with the application systems and initiatives that implement enterprise architecture.]

### 1.5 Organization of this Data Management Plan

[This section describes how this data management plan is organized, and it refers to all other plans that complement the data management plan, such as a related project plan or quality assurance plan.]

## 2.0 DATA ACQUISITION

### 2.1 Data Quality Objectives

[This section describes planning, implementation and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use. It should address areas such

as accuracy, bias, precision, minimum detection limits, and completeness and representativeness of data to be collected and processed.

## **2.2 Data Acquisition**

[This section should include information such as a sampling plan and schedule, locations of observations, sample type and methodology, sampling frequency, sample handling and chain of custody, and sample analysis. It may include all of the sections listed below, as applicable:]

### **2.2.1 Identification of Data**

Data variables and coding conventions  
Naming conventions for data sets  
Data capture and tracking system  
Metadata  
Data collection metadata  
User metadata (metadata for retrieval by those outside the project)

### **2.2.2 Applicable data rights and/or access controls**

[This section encompasses both data security management and data protection; it should also include a description of plans, where needed, for protection of privacy, confidentiality, security, intellectual property and other rights.]

### **2.2.3 Control of Erroneous Data**

### **2.2.4 Changes to Data (Due to Processing or Other Reasons) and Versioning**

[This section documents the planning, implementation and control activities to enable easy access to high quality, integrated data and metadata.]

### **2.2.5 Change Control of the Data**

[This section describes the process for managing change to the data collection or to a data set, including the steps needed to ensure that change is controlled and that its impact assessed before a consensus among partners that the change is needed and beneficial.]

### **2.2.6 Data Validation**

### **2.2.7 Expressing the Amount of a Substance**

Common project parameters and recommended reporting units and formats  
SI multiples, prefixes, and symbols  
Reporting data below minimum detectable limits

### **2.2.8 Time Averaging**

### **2.2.9 Qualifying Data**

## **3.0 DATA MANAGEMENT**

[This section covers data operations and support for structured data assets across the data lifecycle, from creation and acquisition through archival and purge.]

### 3.1 Types of Data to be Collected, Processed, and Utilized

### 3.2 Data Sources

Primary raw data generated from [source, e.g., sensor arrays]

Primary raw data generated from [source, e.g., labs]

Primary raw data requested from third parties

### 3.3 Data Management Resources Needed

[This section describes the resources, such as hardware, software and personnel, needed to accomplish data management for the project. For example: *The Data Management activities will require special computer software, in addition to those for word processing and spreadsheet, to accomplish the goals. SAS (SAS Institute, Cary, NC) will be the default statistical software used. Other computer software that will enable the server computer include, but not limited to, Microsoft SQL Server 2000 (or SQL Server 7.0, a relational database management system) and Windows 2000 Data Center Server (a server operating system, or Windows 2000 Advanced Server). The Data Management Office will take the lead in coordinating the data management activities in the data repository. The Office will be in charge of, with assistance of the data repository QA/QC staff, designing and implementing the overall data management plans.*]

### 3.4 Data Acquisition Activities

[This section describes how data will be collected or generated to support the research project. For example, on a project using field observations gathered by sensor arrays, and lab-generated data, as well as data from third parties, the following might apply: *Data will be received routinely from a number of sources, which are discussed in the following sections. In order to assure that no data are lost, the Data Management Office will utilize a data login procedure to document the receipt of all routinely scheduled data. Separate entries will be created for each of the data sources, and the period of the data received along with the dates the data were received and loaded into the data repository's processing and storage systems. Any gaps in data periods noted during this login sequence will be investigated and resolved immediately.*]

[Example: **Primary Raw Data Generated from the Sensor Array:** *The majority of the primary raw data will be generated from the sensor array. The sensor array contains a variety of aerosol instrumentation. These data will be used to characterize the physical, chemical, and spatial components of particulate air pollutants in the air shed under study. An on-site data acquisition computer, located aboard the sensor array, will be utilized to track and initially process the primary raw data.*]

[Example: **Primary Raw Data Generated from the Lab:** *For those instruments generating sampling mediums (such as air filters) rather than purely electronic data, analyses will be conducted by in-house Lab or other external laboratories, as necessary. All information related to analysis of the samples is considered as primary raw data as well. The analytical results will be transferred to the repository computer, which will be linked with sampling notes associated with original conditions in the sensor array.*]

[Example: **Primary Raw Data Requested from Third Parties:** *Data from third-party sources may be used to enhance the research activities in the project. For example, local traffic counts represent the density of major highways, which will be used to correlate with biological effects of air pollution for In vivo as well as epidemiologic studies. These data will be requested from local traffic authorities and be treated as primary raw data. The other sources of third-party data include those from AQMD, where they have been periodically collected air quality along with meteorological data. These data will be served an external source for crosschecking the validity of the data measured by the sensor array.*]



### 3.5 Data Storage Activities

[This section should document plans for data and metadata content and format, including description of documentation plans and rationale for selection of appropriate standards. Existing, accepted standards should be used where possible. Where standards are missing or inadequate, data sharing strategies and formats for storage and exchange should be developed.]

- **Short-term data storage:**
- **Long-term data storage:**

### 3.6 Data Evaluation and Processing Activities

[This section describes how data will be processed and evaluated for valid use. This is where project planners address areas such as how to code the value of a substance, or how to handle data that are below MDL. Examples of content for this section follow.]

### 3.7 Time Integrated Data

[Example: *Time-integrated data may include times of the beginning and the end of the time averaging period. A valid time-averaged data must contain at least NN% of validated data Points out of the total data points possible for the averaged time period. For example, a 60-minute time-averaged data based on 1-minute samples must contain at least nn validated 1-minute data points. Otherwise, the time-averaged value will be flagged, reported, and potentially considered as missing datum.*]

### 3.8 Date and Time Formats

[Example: *dates and times will be reported for all measurements and in two formats: Pacific Standard Time (PST) and Coordinated Universal Time (UTC, where  $UTC = PST + 8 \text{ hr}$ ). Both the begin time and end time will be reported in both time formats.*]

### 3.9 Reporting Missing Data

[Example: *All data fields should have a value present - either the measured, the adjusted, or a missing value. There should be no blank data fields. Data generators should report data where possible and use flag codes (see 5.5.7 for details). All missing values should be numerical values, not character or alphanumeric values, to aid quality-control efforts. Missing values for data parameters should be represented by a value of -9999.*]

### 3.10 Reporting Calibration Values and Uncertainty Estimates

[Example: *The calibration values, estimates of precision and MDL for all measurements will be maintained by the research investigators and reported to the Data Management Office in separate files other than the main databases. Access to these data is crucial for future quality-assurance, analytical, and modeling exercises. Uncertainty estimates, if available, should be reported. These estimates will be reported either in the measurement method information table.*]

### 3.11 Data Flags

[Example: *Every data record will have an associated data qualification flag code, in addition to any field or laboratory data qualifiers, if applicable. Flags begin with the letter "V" for valid values, "M" for missing values, and "H" for historical data and third-party data that are unable to be assessed or validated by the data repository. Invalid data will not be submitted to the permanent database archive, but will be kept by individual investigators and the analytical lab.*]

### 3.12 Data Access Provisions

[This section is a description of plans for providing access to data. This should include a description and rationale for any restrictions on who may access the data under what conditions over the data’s lifecycle. Note that the data lifecycle is different from the project’s lifecycle. This should also include an estimate of the resources—equipment, connections, systems, expertise, etc.—needed to meet anticipated requests for data. These resources and capabilities should be appropriate for the projected usage, addressing any special requirements such as those associated with streaming video or audio, movement of massive data sets, etc.]

## 4.0 DOCUMENTING DATA

### 4.1 PROJECT XXX Data and Information Categories

[This section describes the types of data that the project will generate and use. For example: *Project XXX will use data and information in four major categories: in situ observations, remote sensing observations, model outputs, and geographic information system (GIS) coverages. The data transmittal format may be different for each of these categories, and data may be stored locally in a number of ways to maximize efficiency. It is important, however, that data transmitted to other Project XXX participants conform to a common set of unit, syntax, and format conventions. Limiting the number of formats will facilitate data and research product sharing and minimize the conversions necessary to load the data into the permanent repository.*]

### 4.2 PROJECT XXX Data Format and Codes

### 4.3 Metadata Tables

### 4.4 Measurement Data Table

[Sections 4.2, 4.3 and 4.4 are similar in that they describe the conventions for encoding values, for data, metadata and measurements. This information can be described in many ways; it is important that, regardless of format, the coding guidance be as complete as possible and a mechanism for adding to or editing the coding values exists. Below is an example of the coding of data values follows, in tabular format, with instruction on how to extend the guidance:]

Chemical species, group, or parameter	Units	Reporting format	Variable name
O <sub>3</sub>	ppbv	xxx.x	o3
NO, NO <sub>y</sub> , or NO <sub>x</sub>	ppbv	xxx.xx	no, noy, or nox
PAN	ppbv	xxx.xx	pan
CO	ppbv	xxx	co
SO <sub>2</sub>	ppbv	xxx.xx	so2
Hydrocarbons	ppbv	xxxx.xx	See Appendix
Carbonyls	ppbv	xxxx.xx	See Appendix
Wind speed	m/s	xx.x	winspd
Wind direction	decimal degrees	xxx.x	windir
Temperature, temperature, or dew point	degrees C	xxx.x	temp, dtemp, and dewpnt
Mixing height	m agl	xxxx	mixhgt
Relative humidity	%	xxx.x	relhum
Solar radiation	Watts/m <sup>2</sup>	xxx.x	solrad
UV radiation	Watts/m <sup>2</sup>	xx.x	uvrad
Pressure or partial pressure	pascals	xx.x	press

Chemical species, group, or parameter	Units	Reporting format	Variable name
Barometric pressure	mb (adjusted to sea level)	xxxx.x	bpress
Precipitation	mm	xxx.x	precip
Altitude	m above msl	xxxxx.x	alt
Latitude and Longitude	decimal degrees	+ - xxx.xxxxx	latdec, londec

If you cannot locate a needed entry in the above table, we recommend the use of standard SI units. If you are still unsure of the proper reporting convention, please contact the data repository for further consultation.

#### 4.5 Data Formatting and Validation Activities

#### 4.6 Data Formatting Activities

#### 4.7 Data Validation Activities

#### 4.8 Reference and Master Data Management

[This section delineates the planning, implementation and control activities that ensure that contextual data is linked to the correct version of data sets. This insures consistency in program and project management documents (and other semi-structured content) and the data sets to which the semi-structured content is related.]

### 5.0 ARCHIVAL AND DISSEMINATION OF PROJECT XXX DATA AND INFORMATION

#### 5.1 Data Archiving Activities

[This section contains a description of plans for preserving data in accessible form. Plans should include a timeline proposing how long the data are to be preserved, outlining any changes in access anticipated during the preservation timeline, and documenting the resources and capabilities—e.g., equipment, connections, systems, expertise—needed to meet the preservation goals. Where data will be preserved beyond the duration of direct project funding, a description of other funding sources or institutional commitments necessary to achieve the long-term preservation and access goals should be provided.]

##### 5.1.1 Transfer of Responsibility

[This section describes plans for changes in preservation and access responsibility. Where responsibility for continuing documentation, annotation, curation, access, and preservation—or its counterparts, de-accessioning or disposal—will move from one entity or institution to another during the anticipated data lifecycle, plans for managing the exchange and documentation of the necessary commitments and agreements should be provided. Data re-use and re-purposing should be described, and agencies should be alerted to needs for standards development or evolution.]

#### 5.2 PROJECT XXX Permanent Database Archive

#### 5.3 Data and Research Product Access and Distribution

## **6.0 RECORDS MANAGEMENT**

### **6.1 Records Management System**

### **6.2 Records Identification, Authentication and Indexing**

### **6.3 Records Distribution and Storage**

### **6.4 Records Retrieval**

### **6.5 Records Retention Reviews**

#### ***6.5.1 Assessing continuing impact or value of the data***

[Discuss the possible impact of the data within the immediate field, in other fields, and any broader, societal impact. Indicate how the data management plan will maximize the value of the data.]

#### ***6.5.2 Determining “Uniqueness” of the Data Collection***

#### ***6.5.3 Risks Associated with Retention***

## **7.0 SOFTWARE CONFIGURATION CONTROL AND DATA MANAGEMENT SYSTEM ADMINISTRATION GUIDELINES**

### **7.1 Project-specific Database and Software Requirements**

### **7.2 Day-to-day Operation of Data Management Systems**



Harnessing the Power of Digital Data: Taking the Next Step March 31, 2011

## APPENDIX E



## APPENDIX E: Agenda for the SDM Workshop



### SCIENTIFIC DATA MANAGEMENT (SDM) FOR GOVERNMENT AGENCIES:

#### WORKSHOP TO IMPROVE SDM

CO-SPONSORED BY CENDI, IWGDD, AND EPA

JUNE 29-JULY 1, 2010

### DAY 1: June 29, 2010

#### 8:15-9:00 REGISTRATION AND MORNING REFRESHMENTS

#### 9:00-10:45 OPENING PLENARY SESSION

- Introduction and welcome  
**Robert Shepanek, EPA**
- Importance of scientific data in decision making and policy development  
**Dr. Pai-Yei Whung, EPA, Chief Scientist, Office of the Science Advisor**
- Importance of Managing Data as a strategic national asset: *"Harnessing the Power of Digital Data for Science and Society"*  
**Chris Greer, Assistant Director, Information Technology R&D for the White House Office of Science and Technology Policy, and Co-chair of the NSTC Interagency Group on Digital Data**
- Results of participant web survey of the current landscape of agency policies and data management planning  
**Bonnie C. Carroll, CENDI, Executive Director**

#### 10:45-11:00 BREAK

#### 11:00-2:45 TECHNICAL PLENARY (Agency presentations on their science data policies, data management planning approaches and challenges) **George Strawn (NITRD)** to moderate

- **National Science Foundation**  
**Phil Bogden, Program Director, Office of Cyberinfrastructure**
- **National Institute of Standards and Technology**  
**Daniel G. Friend, Chief, Thermophysical Properties Division**
- **National Aeronautics and Space Administration**

**Joseph Bredekamp**, *Senior Science Program Executive*

**WORKING LUNCH PROVIDED**

- **Department of the Interior**  
*Karen Siderelis, Geospatial Information Officer*
- **National Institutes of Health/National Library of Medicine**  
*Jerry Sheehan, Assistant Director for Policy Development*
- **National Oceanographic and Atmospheric Administration**  
*Donald Collins, Principal Investigator for the National Environmental Satellite, Data, and Information Service of the National Oceanographic Data Center*

**2:45-3:15**                    **PANEL DISCUSSION**

- Presenters form a panel for cross cutting discussion

**3:15-3:30**                    **CHARGE TO BREAKOUT SESSIONS**

- Description of different types of users and charge to breakout groups  
**H. K. 'Rama' Ramapriyan (NASA)** to moderate
  - Researchers (use data for conducting scientific research; end product – publications, derived data products, and/or models)
  - Science Managers (fund and manage scientific research – e.g., Program Managers in NSF, NASA, etc.)
  - Operational Users (use data on a regular basis in operations. E.g., firefighting, weather forecasting, air quality assessment, crop assessment, topographic mapping)
  - Policy Analysts (use data for making policies – e.g., cap-and-trade, etc.)

**CHARGE 1:** Based on an outline of suggested elements, the group is to 1) ensure the outline accounts for the best practices that should be addressed; 2) draft content for these best practices from the perspective of their communities of practice so that the policy would apply at any science agency; and 3) create a list of issues the group perceives as impediments to implementation of the policy across agencies, and suggest solutions to them.

**3:30-5:00**                    **BREAKOUT SESSIONS BEGIN**

Recommendations for Data Policies - (3 hours total, begin Day 1, end morning of Day 2)

**DAY 2: June 30, 2010**

**9:00-10:30**                    **CONTINUE DATA POLICIES BREAKOUT GROUP DISCUSSIONS**

**10:30-10:45**                    **BREAK**

**10:45-11:45**                    **REPORT TO PLENARY FROM DATA POLICY BREAKOUTS**

**11:45-1:00**                    **CHARGE TO BREAKOUT SESSIONS & BEGIN DISCUSSIONS**

- Recommendations for Data Management Plan Elements - (3 hours total, continued after lunch)

H. K. 'Rama' Ramapriyan (NASA) to moderate

- Researchers
- Science Managers
- Operational Users
- Policy Analysts

**CHARGE 2:** Based on the preliminary list of elements for a data management plan that will be provided, the group is to 1) ensure the outline accounts for all the elements that should be addressed; 2) develop descriptions of the elements from the perspective of their communities of practice such that any agency could develop its data management plans based on this outline; and 3) create a list of issues the group perceives as impediments to implementation of the elements of the plan at their agencies, or within their communities of practice and suggest solutions to them.

**WORKING LUNCH PROVIDED**

**1:30-3:30 CONTINUE DATA MANAGEMENT PLAN BREAKOUT GROUP DISCUSSIONS**

**3:30-3:45 BREAK**

**3:45-4:45 REPORT TO PLENARY FROM DATA MANAGEMENT PLANS BREAKOUTS**

**DAY 3: July 1, 2010**

**9:00-10:00 CLOSING PLENARY**

- Panel of Reactors to Breakout Group Recommendations  
**Representatives of the sponsoring organizations:**
  - **Chuck Romine**, *Senior Policy Analyst, Office of Science and Technology Policy, Executive Office of the President*
  - **Jerry Blancato**, *Director, EPA Office of Research and Development, Office of Administration and Research Support*
  - **Donald Hagen**, *Associate Director, Office of Product Management and Acquisition, National Technical Information Service for the U.S. Department of Commerce*

**10:00-12:00 FINAL GENERAL DISCUSSION**

- Formulation of Recommendations and Action Plan
- GEOSS as Case Study
  - **Gary Foley**, *Senior Advisor to the EPA Chief Scientist*

**12:00 CLOSE OF MEETING**

**12:00-4:00 DRAFTING THE WORKSHOP REPORT, Volunteers Welcome**

**PRODUCT OF THE WORKSHOP**

- Report to the sponsors with a set of recommendations on best practices for data management policies and plans.